

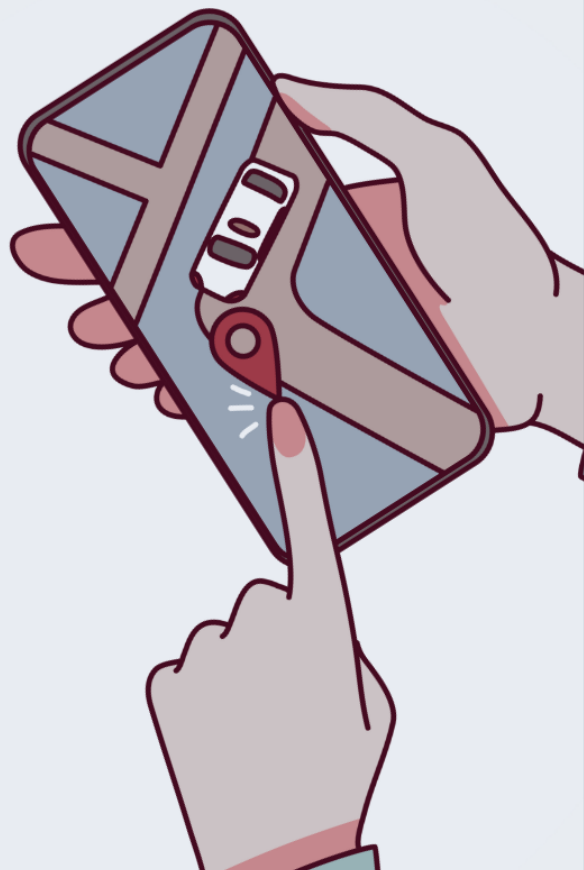
# End-to-End Business Intelligence Solution



**Ride4all**

## **Group 40**

Catarina Nunes 20230083  
Francisco Castro 20230992  
Pedro Catarro 20230463



## **Index of contents**

1	Introduction .....	5
2	The Organization.....	6
2.1	Presentation of the organization .....	6
2.2	Problem Scenario: Business Needs.....	8
2.3	Business Questions .....	9
3	Original Data Sources .....	11
3.1	Data Source.....	11
3.2	Dataset Structure .....	11
4	Data Warehouse.....	12
4.1	Dimension Tables of the Data Warehouse .....	13
4.2	Hierarchies .....	18
4.3	Fact Table of the Data Warehouse: Fact_Rides .....	19
4.4	Star Schema Model.....	20
5	ETL Process.....	21
5.1	Identity Data Issues .....	22
5.2	Loading the Data into the Data Warehouse.....	23
	PL_Ride4All_LOAD_DW .....	23
	PL_Ride4All_VALIDATE_RULES.....	29
	PL_Ride4All_CHECK_FAILS.....	30
6	Semantic Model .....	32
6.1	Connection to Power BI Desktop .....	32
6.2	Model optimization.....	33
6.3	Measures and calculated columns .....	36
7	Power BI report.....	39
7.1	2023 Overview .....	40
7.2	Performance Analysis and Forecast.....	41
7.3	Driver Analysis (Driver Profile).....	42
7.4	Client Analysis (Client Profile).....	43
7.5	Location Analysis.....	44
7.6	Operational & Car Analysis .....	45
8	Extra: Row Level Security.....	47

*Nova Information Management School, Lisbon, Portugal*

9	Learning Curve & Critical Review .....	49
9.1	Date, Time & Location Dimensions .....	49
9.2	Period of the Day Definition .....	49
9.3	Map Visualizations .....	50
9.4	Constant Ratios .....	50
10	Conclusion: Main aspects & Key takeaways .....	51
11	Appendix: Executed procedures .....	52
1.	Duplicate_Locations_to_map .....	52
3.	log_quality_checks: Validation Rules and Row Count. ....	53
4.	Email Notification:.....	53
5.	Power BI Report .....	54
	2023 Overview .....	54
	Performance Analysis & Forecast .....	55
	Driver Analysis.....	56
	Client Analysis .....	56
	Location Analysis .....	57
	Operational Analysis & Car Analysis .....	57

## Index of Figures

<b>Figure 1.</b>	Growth of Ride4all's Total Rides Over the Years (Excl 2023 YTD).....	6
<b>Figure 2.</b>	Growth of Ride4all's Total Rides Over the Years: Quarterly Analysis ....	6
<b>Figure 3.</b>	Coverage Map of Ride4all Across Portugal .....	7
<b>Figure 4.</b>	Ride4all vehicles fleet distribution.....	7
<b>Figure 5.</b>	Relationship Editor .....	17
<b>Figure 6.</b>	Star Schema Model .....	20
<b>Figure 7.</b>	Lakehouse Ride4All_LH_SOURCES.....	21
<b>Figure 8.</b>	Script Activity.....	23
<b>Figure 9.</b>	Wait Activity .....	23
<b>Figure 10.</b>	Dataflow_Car.....	23
<b>Figure 11.</b>	Dataflow_Client.....	24
<b>Figure 12.</b>	Dataflow_Currency .....	24
<b>Figure 13.</b>	Dataflow_Date .....	25
<b>Figure 14.</b>	Dataflow_Drivers.....	25
<b>Figure 15.</b>	Dataflow_Location.....	27
<b>Figure 16.</b>	Dataflow_Time .....	27
<b>Figure 17.</b>	Dataflow_Fact_Rides .....	29
<b>Figure 18.</b>	PL_Ride4All_VALIDATE_RULES.....	30

<b>Figure 19.</b> PL_Ride4All_CHECK_FAILS.....	31
<b>Figure 20.</b> PL_Ride4All_LOAD_DW .....	31
<b>Figure 21.</b> D Date Optimizations.....	34
<b>Figure 22.</b> D Time Optimizations .....	34
<b>Figure 23.</b> D Dropoff Location and D Pickup Location Optimizations .....	35
<b>Figure 24.</b> D Car Optimizations.....	35
<b>Figure 25.</b> Final Semantic Model .....	38
<b>Figure 26.</b> Daily Distribution of Rides (Aveiro, Q1 2023) .....	41
<b>Figure 27.</b> Driver Utilization Rate of Total Kilometers Driven.....	47
<b>Figure 28.</b> Row Level Security filter for Drivers.....	48
<b>Figure 29.</b> Row Level Security Example.....	48
<b>Figure 30.</b> Duplicate_Locations_to_map table.....	52
<b>Figure 31.</b> Successful pipeline running.....	53
<b>Figure 32.</b> Results from the rules.....	53
<b>Figure 33.</b> Email of Successful pipeline running.....	53
<b>Figure 34.</b> Report Cover.....	54
<b>Figure 35.</b> 2023 Overview .....	54
<b>Figure 36.</b> Performance Analysis.....	55
<b>Figure 37.</b> Forecast .....	55
<b>Figure 38.</b> Driver Profile .....	56
<b>Figure 39.</b> Driver's Routes Analysis .....	56
<b>Figure 40.</b> Client Profile.....	56
<b>Figure 41.</b> Location Analysis .....	57
<b>Figure 42.</b> Operational Analysis .....	57
<b>Figure 43.</b> Car Analysis .....	57

## Index of Tables

<b>Table 1.</b> Business Questions and Measures .....	9
<b>Table 2.</b> Client Dimension.....	13
<b>Table 3.</b> Car Dimension .....	13
<b>Table 4.</b> Driver Dimension .....	14
<b>Table 5.</b> Location Dimension .....	15
<b>Table 6.</b> Date Dimension.....	16
<b>Table 7.</b> Time Dimension.....	16
<b>Table 8.</b> Currency Dimension.....	17
<b>Table 9.</b> Rides Fact Table.....	19
<b>Table 10.</b> Data Quality Issues found using the Notebook_EDA.....	22
<b>Table 11.</b> Validation Rules .....	30
<b>Table 12.</b> Measures .....	37
<b>Table 13.</b> Calculated columns.....	37

*Nova Information Management School, Lisbon, Portugal*

## **1 Introduction**

In the current business landscape, robust data management tools are no longer considered just an advantage, but rather an essential component for enhancing decision-making across all organizational levels. As companies expand, their solutions must grow in tandem to support their growth. Data can offer a competitive advantage to those who understand its value.

Our project aligns with this idea by utilizing a Business Intelligence solution in the form of a Data Warehouse to optimize decision-making processes and position businesses strategically for greater success in the market. This approach views data not merely as a resource, but as a powerful tool for transforming businesses. Nowadays, it is increasingly necessary for companies to know how to structure and analyze their data correctly to ensure their position in the market, as is the case with Ride4ALL.

This is a project that aims to assist this ride-sharing company in improving its analytical output and decision-making process using business intelligence methodology.

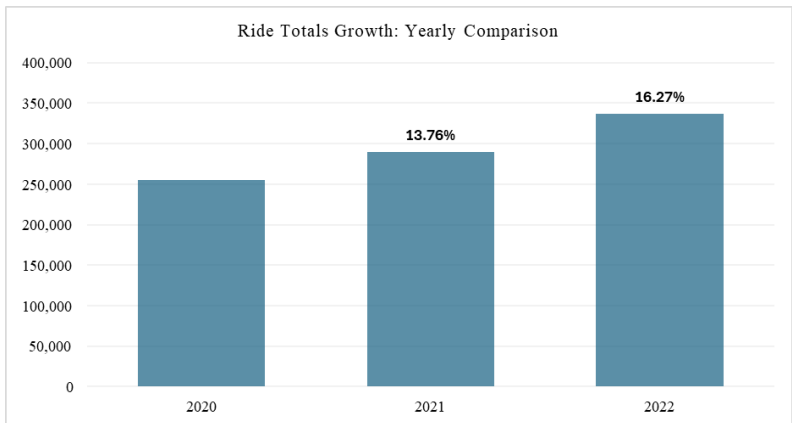
## 2 The Organization

### 2.1 Presentation of the organization

Ride4all was founded in 2019 with a goal to make transportation more accessible and convenient in Portugal. The company started as a startup, emphasizing innovation, efficiency, and customer satisfaction. The company provides a platform where passengers can connect with drivers through their app.

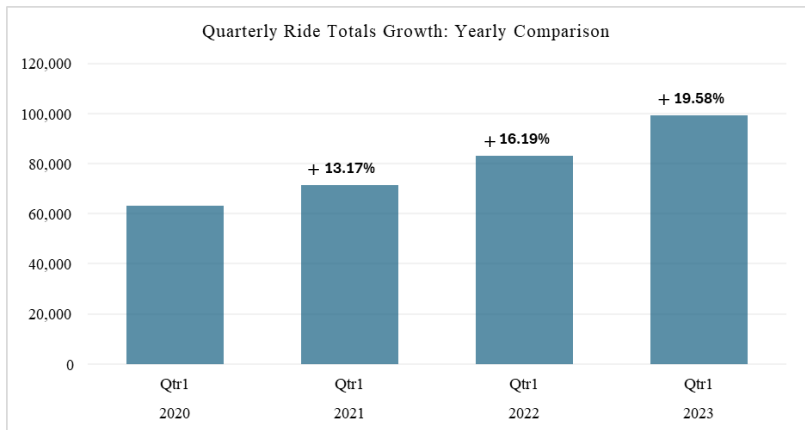
Ride4all distinguishes itself from competitors like Uber or Bolt with its standout safety features: enabling customers to create profiles and select preferred drivers, nurturing familiarity, and trust. This can lead to having the same customer with the same driver on multiple rides.

Despite facing initial challenges due to the pandemic, Ride4all has grown in the Portuguese ride-sharing market over the past three years:



*Figure 1. Growth of Ride4all's Total Rides Over the Years (Excluding 2023 YTD)*

In early 2023, Ride4all's administration witnessed significant growth and stability in the company's operations:

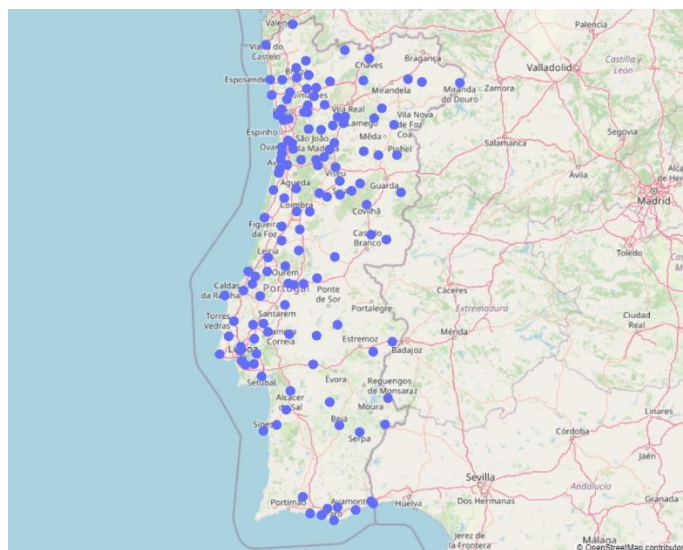


*Figure 2. Growth of Ride4all's Total Rides Over the Years: Quarterly Analysis*

*Nova Information Management School, Lisbon, Portugal*

Recognizing this milestone, the administration made the strategic decision to invest in more advanced data-driven tools.

Furthermore, with a clientele of 10k users and Ride4all's expansion across Portugal, the need to enhance its network management also grew:

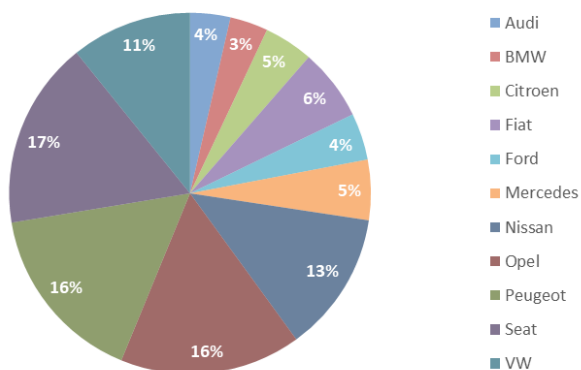


*Figure 3. Coverage Map of Ride4all Across Portugal*

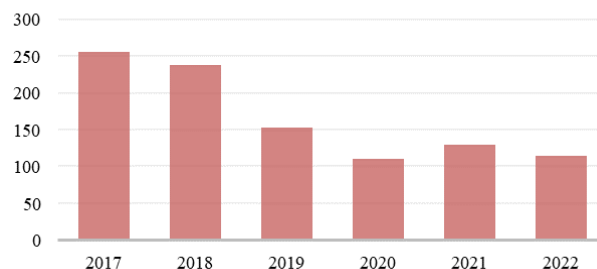
Furthermore, Ride4all offers unique income opportunities for families. For example, a couple can jointly own and operate a vehicle, taking turns driving and earning. This innovative model not only provides financial support for families but also offers flexibility for drivers.

With a fleet comprising 500 recent vehicles and a team of 1000 professional drivers, Ride4all has solidified its position as a trusted transportation partner within the industry.

Car Distribution by Manufacturer



Car Distribution by Manufacturing Year



*Figure 4. Ride4all vehicles fleet distribution by Manufacturer (left) and by Manufacturing Year (right)*

*Nova Information Management School, Lisbon, Portugal*

## 2.2 Problem Scenario: Business Needs

Facing the company's growth trajectory, the administration acknowledges the pivotal role of data in driving informed decision-making and the need to improve analytical capabilities. Furthermore, Ride4all faces the challenge of using the large amount of information available in their daily refreshed server data to achieve updated analysis and faster decision-making. They aim to understand patterns, trends, and customer preferences to make their services better, work more efficiently, and stay competitive in the ride-sharing market.

To deal with this challenge, Ride4all plans to establish a central data repository as the foundation of its analytical infrastructure. By bringing together data from different sources, Ride4all hopes to make its analytical processes simpler, integrate data better, and give smooth access to important information. The central approach makes sure that the data is accurate and consistent and prepares for future expansions and operational improvements.

This is where our team comes in. Ride4all is looking for a Business Intelligence solution that can help them analyze specific things:

- **Trips per month, Quarter, and Yearly (accumulated to date):** By accumulating this data over time, they want to analyze trends in ride frequency and identify peak periods of demand.
- **Number of Rides as well as Average Distance of Rides:** They want to understand the overall service utilization and assess the efficiency of their transportation network.
- **Comparison of current year vs previous year:** With this analysis, they want to gauge growth trajectories, identify areas of improvement, and track progress over time.
- **Building client profiles:** To better understand customer preferences and travel habits so it is possible to tailor their services and customer experience effectively.
- **Ride details (at the trip level):** They need detailed information for each ride, including start and end times, pickup and drop-off locations, driver details, and fare amounts. This ride-level analysis allows granular knowledge of individual trip dynamics.
- **Aggregate Rides by driver, pick-up point, and destination point:** To identify top-performing drivers, high-demand areas, and popular destination routes, informing resource allocation and operational decisions.



*Nova Information Management School, Lisbon, Portugal*

## 2.3 Business Questions

After several meetings with the team responsible for rides management, and based on the specific requirements listed previously, we identified six major categories to guide our questions: Overall Performance, Driver Performance, Client Analysis, Location Analysis, Operational Metrics, and Vehicle Analysis. The result is the following set of business questions that can evolve in future discussions with the client:

*Table 1. Business Questions and Measures*

Category	Business Question
<b>Overview</b>	Who are the top 5 drivers in terms of total rides in 2023?
	What is the total number of drivers who have completed rides Year-to-Date (YTD) in 2023?
	What are the top 5 performing locations in terms of the number of rides in 2023?
	What are the most demanded hour ranges of the day based on the number of rides and the start time of the ride, in 2023?
	What is the average fare per trip for the top driver in 2023?
	What is the peak hour (in terms of the number of rides) at the top pick-up District (in terms of the number of rides) during the first quarter of 2023?
<b>Performance Analysis &amp; Forecast</b>	Who was the best driver in terms of total rides in 2022?
	What is the most demanded hour of the day on the busiest day of the week based on the number of rides and the start time of the ride across the years (2020-2022)?
	What is the average number of rides percentage variation between weekdays and weekends in the first quarter of 2023?
	What is the average duration of rides in the busiest month of 2022?
	What is the total projection (forecast for the number of rides) for 2023?
<b>Driver Analysis</b>	<b>Choosing on specific driver:</b>
	What is the total number of rides (all years)?
	What is the period of the day with the most rides on the day of the week with the highest percentage of rides (all years)?
	What is the average ride distance in kilometers in the last month of analysis (May 2023)?
	In the pick-up district with the most rides (all years), what is the most repeated route?

*Nova Information Management School, Lisbon, Portugal*

<b>Client Analysis</b>	<b>Choosing on specific client:</b>
	Who are the drivers with the most repeated rides from this customer (all years)?
	What is the route and the driver of the most expensive trip taken by the customer?
	On the day of the week with the most rides by the client, what is the hour with the most rides and what is the average ride duration of those rides (all years)?
<b>Location Analysis</b>	What is the period of the day with the most rides in the top pick-up location?
	Considering the top pick-up location (district), what is the top drop-off location (district) with the highest number of rides and which day of the week has the highest demand for this route?
	Within the combination from the previous question, what is the route (by cities) with the highest number of rides and what is its average fare per trip?
	What is the forecast in terms of the number of rides for the next full month (06/2023) for the route mentioned in the previous question?
<b>Operational Analysis &amp; Car Analysis</b>	What are the top 5 manufacturers in terms of number of rides?
	What is the average gas consumption per kilometer and the average gas consumption per amount for the manufacturer with the most rides?
	What is the manufacturer with the highest average gas consumption per km?
	What are the top 3 cars with the highest number of rides?
	What are the top 3 cars with the lowest number of rides?
	In the car with the most rides, what is the usage rate of kilometers for each driver?

To answer the business questions, we need to ensure that our fact table contains the necessary measures and dimensions. The existing measures (Duration of each ride, Distance (in Kms) traveled, Amount earned, and Gas consumption) are already present in the file. Additionally, we can enhance the table by including pre-calculated ratios like *Amount/Km*, *Duration/Km*, *Gas/Km*, and *Gas/Amount* for better analysis.

While some questions require the **count of rides** meeting specific conditions, our fact table is structured to represent **individual rides**. Therefore, we don't need the "number of rides" as a direct measure. Instead, we can aggregate rows based on conditions, allowing us the flexibility to analyze data based on the required criteria.

*Nova Information Management School, Lisbon, Portugal*

### 3 Original Data Sources

#### 3.1 Data Source

Data for the project was provided by Ride4all's administration. CSV files were extracted directly from the company's main servers, with the server refreshed daily to incorporate new data. The historical data spans approximately 3 years, from 2020 to 2023 (June).

#### 3.2 Dataset Structure

- **Cars and Drivers A.csv and Cars and Drivers B.csv:** These files provide data on vehicles and drivers. Cars and Drivers A.csv represents female drivers, while Cars and Drivers B.csv represents male drivers, the same car having both a female and a male driver. During the discovery process, it was observed that female drivers have significantly fewer drives compared to their male counterparts.
- **customers.csv:** This file contains information about the customers who use the ride-sharing service, including registration details.
- **Locations.csv:** Provides data on geographical locations relevant to the ride-sharing service such as city and country.
- **Rides.csv:** Contains detailed information on each ride transaction, including, timestamps, duration, driver, pickup/drop-off locations, Gas, Kms, and Amounts.

During the exploration of this source file, it was noted that some ride transactions have a common amount value of 3.6. This common fare amount may indicate minimum fees or cancellation fees, but confirmation from Ride4all is required to interpret its significance. Additionally, the maximum duration of the rides is 6060 seconds.

Amount	Count of Ride ID
3.6	170085

Furthermore, this file also provides insights into the dimensional model required for the company. One note is the existence of both start and end date-time, pick-up location, and drop-off location, which will require attention during the design phase.

*Nova Information Management School, Lisbon, Portugal*

## 4 Data Warehouse

After discussing the best approach for our design, we reached the conclusion that following the **Kimball methodology** was the way to start. This methodology offers several advantages, including proven effectiveness, relatively quick setup, and low initial costs. Additionally, it provides simplicity, efficient query performance (especially for querying large datasets), and scalability, making it suitable for growing databases.

One of the key aspects of the Kimball methodology is its ability to facilitate the creation of a centralized data repository known as the data warehouse. This data warehouse serves as a singular source of truth, integrating and storing data from multiple operational systems in a consistent and structured manner.

A significant technique used in the Kimball methodology is denormalization, which involves including additional descriptive columns beyond the keys. This leads to a **star schema design** that consists of a central fact table surrounded by dimensional tables.

Following this methodology, the first step was to **identify the underlying business process**. Based on the business requirements mentioned in the previous sections, it was clear that **Ride** management was the most crucial business process that had to be represented in our Data Warehouse.

The next step was to **define the granularity**. The grain of the fact table is at the level of individual rides (unique ride transactions). Each record in the fact table represents a single ride taken by a client and includes details such as the driver, car, pickup location, drop-off location, start and end dates and times, and currency used for payment:

***Client x Driver x Car x Pickup Location x Dropoff Location x Start Day x End Day  
 x Start Time x End Time (HH: MM) x Currency***

So each record in the fact table represents a single ride instance, regardless of whether the same client takes multiple rides on the same day. By capturing unique ride transactions, our Data Warehouse provides the flexibility to aggregate data at higher levels such as daily, weekly, or monthly summaries.

This approach enables a detailed view of ride performance across multiple dimensions, facilitating analysis and reporting at different levels of aggregation. Additionally, it preserves the ability to drill down to individual ride instances when necessary, allowing for detailed examination of specific ride characteristics and patterns, just as required by the client.

#### 4.1 Dimension Tables of the Data Warehouse

In the two final phases, we will identify the dimensions (to provide descriptive context) and the facts (containing the measures). We'll provide more details on this process in the following sections.

To maintain data integrity and cohesion in our data warehouse, we use three types of keys: Surrogate Keys, Business Keys, and Foreign Keys. Surrogate Keys are system-generated unique identifiers. Business Keys, in contrast, are natural identifiers derived from real-world attributes, providing identification within the business context, e.g. IDs. Foreign Keys connect tables.

- **Dim\_Client**

Column Name	Data Type	Description
<b>SK_Client</b>	INT	Surrogate Key: Used as a unique identifier for Dim_Client
BK_Client	INT	Business Key: Customer ID
Client_Name	VARCHAR(100)	Name of the client
Client_Birthdate	DATE	Client's Birthday
Client_Gender	VARCHAR(10)	Client's gender. 'Female' or 'Male'

*Table 2. Client Dimension*

- **Dim\_Car**

Column Name	Data Type	Description
<b>SK_Car</b>	INT	Surrogate Key: Used as a unique identifier for Dim_CAR
BK_Car	INT	Business Key: Car ID
Car_Plate	VARCHAR(10)	The car plate, for example: AZ-12-YY.
Brand	VARCHAR(20)	The brand of the car.
Manufacturer	VARCHAR(20)	The Manufacturer who made the car.
Car_Year	INT	Year the car was first registered.

*Table 3. Car Dimension*

Nova Information Management School, Lisbon, Portugal

- **Dim\_Driver**

Column Name	Data Type	Description
<b>SK_Driver</b>	INT	Surrogate Key: Used as a unique identifier for Dim_Driver
BK_Driver	INT	Business Key: Driver ID
Driver_Name	VARCHAR(100)	The name of the driver.
Driver_Birthdate	DATE	Driver's birthday
Driver_Gender	VARCHAR(10)	Driver's gender. 'Female' or 'Male'
Current_Car	INT	The ID of the car the driver is currently driving.

*Table 4. Driver Dimension*

We decided to separate the Driver table and the Car table for two main reasons:

1. **Efficient Storage:** Initially, we had separate files for female and male drivers, but we wanted to consolidate all drivers into a single dimension. However, since each car is assigned to a female driver and a male driver, duplicating car information for each driver would have been wasteful in terms of storage. To optimize storage, we retained only one column to represent the current car of each driver to keep track. Meanwhile, detailed car-related information is stored in a separate dimension, eliminating redundancy, and providing the necessary context for each car.
2. **Granularity Maintenance:** By maintaining the original granularity level of the data in the source file, we ensure that both riders and cars remain at the detail level for each ride. This simplifies queries as we don't need to reference the client table to retrieve car-related information. Having both riders and cars at the same level of granularity facilitates query execution, particularly for our business questions related to cars.

- **Dim\_Location**

Column Name	Data Type	Description
<b>SK_Location</b>	INT	Surrogate Key: Used as a unique identifier for Dim_Location
BK_Location	INT	Business Key: Location ID

*Nova Information Management School, Lisbon, Portugal*

City	VARCHAR(50)	The name of the city.
District	VARCHAR(20)	The name of the district.
Country	VARCHAR(20)	The name of the country.

*Table 5. Location Dimension*

- **Dim\_Date**

Column Name	Data Type	Description
<b>SK_Date</b>	INT	Surrogate Key: Used as a unique identifier for Dim_Date
Proper_Date	DATE	The date. 07/04/2021
Full_Date	VARCHAR(50)	E.g. Wednesday, 1 of January of 2020
Day	INT	7
Month	INT	4
Year	INT	2021
Week	INT	5
Weekday_Number	INT	The number of the day of the week, e.g.: 3.
Weekday_Name	VARCHAR(10)	Day of the week, e.g. Wednesday.
Weekday_Name_Short	VARCHAR(5)	The shot name of the day of the week, for example: Wed.
Weekday_type	VARCHAR(15)	If it's a weekday it takes the value of 'Weekday', 'Weekend' otherwise.
Is_Special	VARCHAR(20)	If that is a special day, it takes that day as value. (Christmas and New Year)
Month_Name	VARCHAR(10)	The name of the month, e.g.: April.
Month_Name_Short	VARCHAR(5)	The short name of the month, e.g.: Apr.
Quarter	INT	E.g. 2

*Nova Information Management School, Lisbon, Portugal*

Quarter_Short	VARCHAR(5)	The short name of the quarter, e.g.: Q2
Quarter_Name	VARCHAR(10)	The name of the quarter, e.g.: Quarter 2
Semester	INT	E.g. 1
Semester_Short	VARCHAR(5)	The short name of the semester, e.g.: S1.
Semester_Name	VARCHAR(10)	The name of the semester, e.g.: Semester 1.

*Table 6. Date Dimension*

- **Dim\_Time**

Column Name	Data Type	Description
SK_Time	INT	Surrogate Key: Used as a unique identifier for Dim_Time
Proper_Time	VARCHAR(10)	4:37 PM
Hour	INT	The hour, for example: 16
Minute	INT	The minute, for example: 37
Hour_Range	VARCHAR(15)	The hour range of the trip, for example: if the rides start at 16:37, the range will be 4 pm -5 pm.
Period_of_Day	VARCHAR(10)	The period of the day: Morning, afternoon, and evening.

*Table 7. Time Dimension*

For the time dimension, we also add the period of the day and time range for analysis purposes.

These last dimensions (Location, Date and Time) demanded special consideration and research/benchmarking to identify industry best practices for this design. After thorough analysis, we narrowed down our options to two main approaches:

- **Creating Multiple Dimensions:** We create multiple dimensions to organize date/time/location-related aspects such as Start date and End date, streamlining data navigation. However, managing multiple dimensions demands careful maintenance, increases storage requirements, and may complicate the schema.



*Nova Information Management School, Lisbon, Portugal*

- **Utilizing Multiple Relationships with USERELATIONSHIP:** This approach involves maintaining only one active relationship. It conserves storage space by retaining a single dimension while accommodating multiple columns in the fact table. However, it necessitates understanding DAX functions, potentially adding complexity. Inactive relationships might confuse users and require careful management.

Although we initially planned our solution according to the first approach, we realized during the ETL process that the second approach would be more advantageous in terms of loading times and storage. This involved having only one dimension of each connected twice, with one relation active and the other inactive. We have updated the final design to reflect this change, and we will manage these relations along the way as necessary, given the end-to-end nature of our project.

☐ Make this relationship active

*Figure 5. Relationship Editor*

- **Dim\_Currency**

Column Name	Data Type	Description
SK_Currency	INT	Surrogate Key: Used as a unique identifier for Dim_Currency.
Currency_Code	VARCHAR(5)	The currency code, e.g.: EUR.
Currency_Name	VARCHAR(10)	The name of the currency, e.g.: Euro.

*Table 8. Currency Dimension*

The Dim\_Currency dimension will contain the names and codes of all currencies used for payment. This dimension will have a positive impact on future analyses of the company. If Ride4ALL decides to expand its market, it is important to know the currency used for payment, especially if it's not the Euro, to accurately determine the true value of the trip using the corresponding conversion rate. Even though this dimension has only one value right now, it is still seen as a level of detail in our model. (It may be an oversimplification, but it's like having only one client; we are still going to create a dimension for its information.)

## 4.2 Hierarchies

- **Date hierarchy**

The **Dim\_Start\_Date** and **Dim\_End\_Date** include year, semester, quarter, month, and day for a better understanding of time, from high-level seasonal trends to more detailed views.



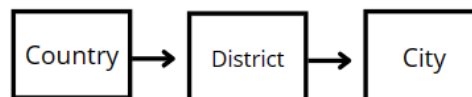
- **Time hierarchy**

The hierarchy below is present in **Dim\_Start\_Time** and **Dim\_End\_Time**. It is good to have this hierarchy because it gives the possibility to understand the demand behavior throughout the day to detect rush Hours and try to reinforce our position during those.



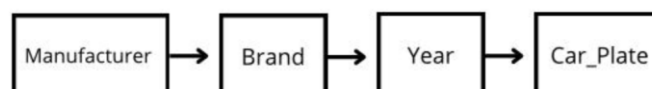
- **Location hierarchy**

The **Dim\_Pickup\_Location** and **Dim\_Dropoff\_Location** incorporate the country, District, and city, where the rides occur. Here we can see broadly or very specifically where customers have been picked up and where they have been dropped off.



- **Car hierarchy**

The **Dim\_Car** hierarchy is structured as follows, almost resembling our "product" hierarchy. We start with a range of manufacturing years, within which there are multiple manufacturers. Each manufacturer can produce multiple brands, and within each brand, we find our car models identified by their unique car plates ("product name").



*Nova Information Management School, Lisbon, Portugal*

#### 4.3 Fact Table of the Data Warehouse: Fact\_Rides

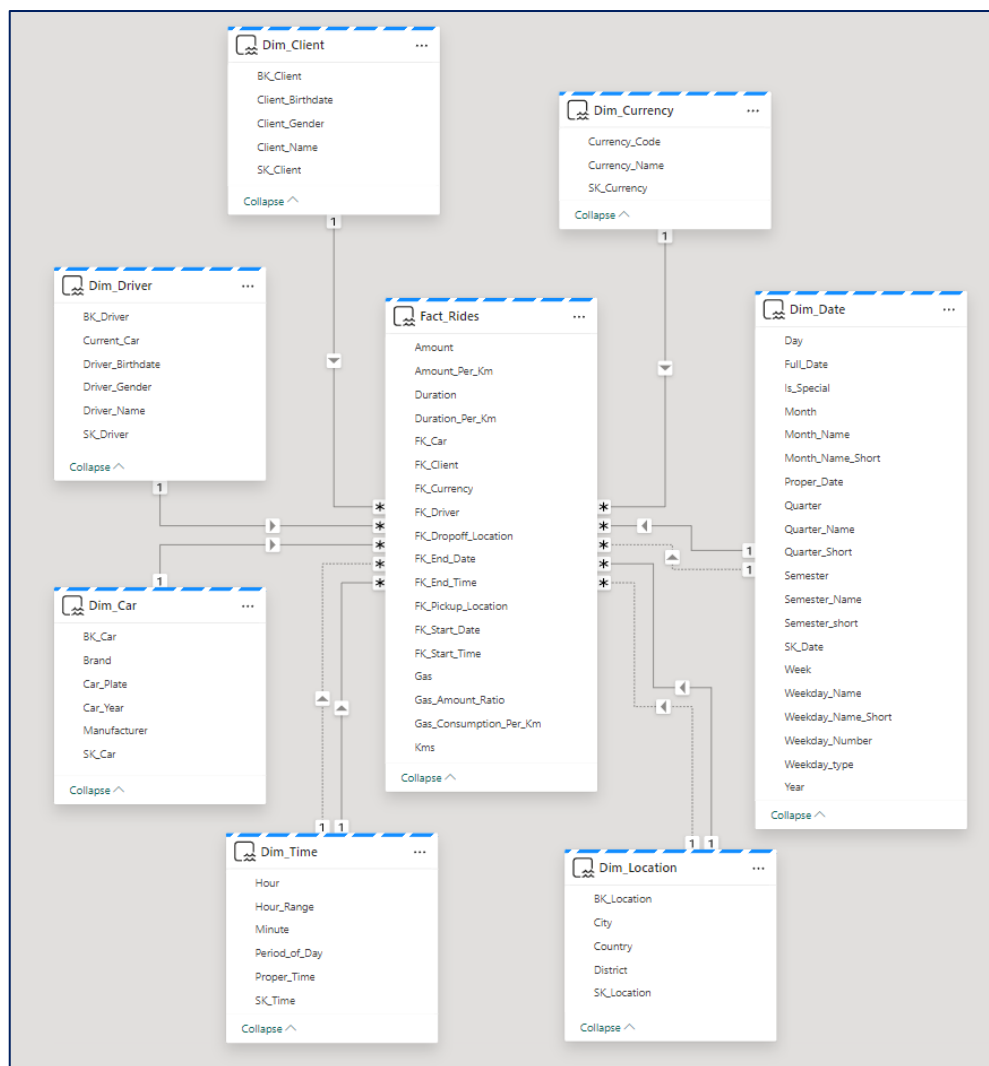
We identified rides as the fact table in our Data warehouse design, that fact table is called **Fact\_Rides** and represents the central table in our Star Schema.

*Table 9. Rides Fact Table*

Column Name	Data Type	Description
FK_Start_Date	INT	Foreign key of Dim_Start_Date.
FK_End_Date	INT	Foreign key of Dim_End_Date.
FK_Start_Time	INT	Foreign key of Dim_Start_Time.
FK_End_Time	INT	Foreign key of Dim_End_Time.
FK_Pickup_Location	INT	Foreign key of Dim_Pickup_Location.
FK_Dropoff_Location	INT	Foreign key of Dim_Dropoff_Location.
FK_Car	INT	Foreign key of Dim_Car.
FK_Driver	INT	Foreign key of Dim_Driver.
FK_Client	INT	Foreign key of Dim_Customer.
FK_Currency	INT	Foreign key of Dim_Currency.
<b>Duration</b>	INT	The duration of the ride (minutes).
<b>Gas</b>	DECIMAL (11, 2)	The amount of gas spent on that ride (Liters).
<b>Kms</b>	DECIMAL (11, 2)	The distance in kilometers of the ride.
<b>Amount</b>	DECIMAL (12, 2)	Price paid by the Customer for the ride.

We are going to convert the duration (currently in seconds) to minutes during the ETL process. The DECIMAL data type of the measures may need adjustments to accommodate the data after the ETL, especially considering some unconventional values in those columns (e.g., kilometers = 7000003)

#### 4.4 Star Schema Model



*Figure 6. Star Schema Model*

Following our initial sketch, the DW is now set up in the Fabric interface with connections established. The Star Schema is in place, with each table featuring a surrogate key (SK) and forming active one-to-many relationships with the fact table's foreign key (FK). Notably, the Date, Time, and Location Dimensions are connected twice, once as active and once as inactive relationships.

*Nova Information Management School, Lisbon, Portugal*

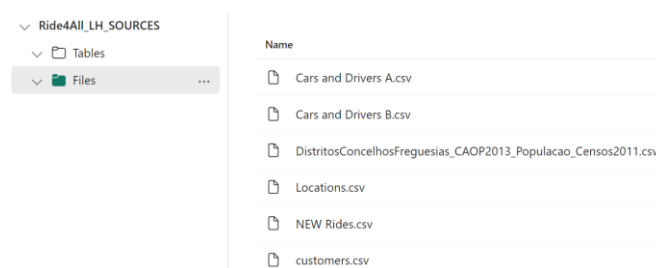
## 5 ETL Process

The next procedure involves selecting data sources for extraction, transforming the data, and loading the transformed data into the Data Warehouse (ETL).

For the ETL process, we utilized the following distinct sources<sup>1</sup>:

- CSV files provided by the company.
- "DistritosConcelhosFreguesias\_CAOP2013\_Populacao\_Censos2011.csv," support file containing information about Portuguese Freguesias, cities, and districts.

During this process, we established a Source Lakehouse named "**Ride4All\_LH\_SOURCES**" where we uploaded our data:



*Figure 7. Lakehouse Ride4All\_LH\_SOURCES*

Additionally, we structured our workflow into the following pipeline arrangement:

- **PL\_Ride4All\_LOAD\_DW:** This pipeline serves as the heart of our solution, executing the ETL process for each dimension and fact table, and subsequently loading the data into the Data Warehouse. Following the completion of the loading process, we have two additional pipelines to aid in verifying that everything proceeded as intended within this pipeline.

The supplementary pipelines are as follows:

- **PL\_Ride4All\_VALIDATE\_RULES:** Conducts quality checks on the tables within the Data Warehouse subsequent to the loading process.
- **PL\_Ride4All\_CHECK\_FAILS:** If any specific rules from the preceding pipeline return a FAIL status, this pipeline is triggered to detect the issue, prompting us to investigate further via email notification.

<sup>1</sup> For Special characters visualization we saved the files from the original Excel sheets in CSV format with utf-8 encoding.

*Nova Information Management School, Lisbon, Portugal*

## 5.1 Identity Data Issues

	Notebook_EDA	Notebook
---	--------------	----------

In order to understand our data and detect errors, we use the Microsoft Fabric notebook tool for Data Exploratory Analysis (EDA). However, this activity was not integrated into the pipeline since it doesn't need to run every time we load the DW. Nonetheless, it's crucial for better understanding of the data before go into the ETL process, so it's available in the solution workspace. Therefore, we identified the following issues:

Source	Issues Found
<b>Cars and Drivers A.csv and Cars and Drivers B.csv</b>	Information about the car and driver are together and we need it separately;
	Information about female and male drivers are in different datasets;
	There are spaces in the beginning of the column title. We need to correct this otherwise we are not able to publish the dataflows;
	Cars and Drivers A has a lot of quotation marks, which prevents the correct recognition of column separations;
<b>Customers.csv</b>	Does not have headers;
	Has 10 duplicates on the client with id 10000;
	Gender info it's not standardize;
<b>Locations.csv</b>	Does not have headers;
	The column with district data has a lot of missing values;
	There are two rows with the same location (Beja) and different id;
	There are two rows with the same location but written differently: 'São João de Madeira' and 'São João da Madeira'; however, 'São João de Madeira' does not exist;
<b>NEW Rides.csv</b>	There is one missing value in the car column;
	We need to separate the date and time into 2 columns;
	After correcting the duplicate values, we need to map the IDs in the rides data and correct them accordingly;

*Table 10. Data Quality Issues found using the Notebook\_EDA*

Nova Information Management School, Lisbon, Portugal

## 5.2 Loading the Data into the Data Warehouse

### PL\_Ride4All\_LOAD\_DW Pipeline



Within our core pipeline, we are using the following activities:

1. **SQL script activities** to clear the current data in the Dimensions and Fact Tables:



Figure 8. Script Activity

It's essential to undergo this phase to ensure a clean slate with every loading. This approach ensures that future loadings with additional entries don't introduce duplicate records, thereby maintaining data integrity.

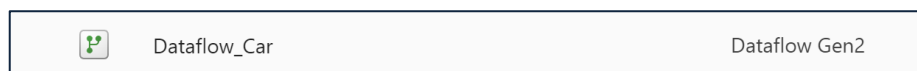
2. **Wait Activities:**



Figure 9. Wait Activity

These activities serve as placeholders. Their inclusion is crucial to ensure the coordination/synchronizing of steps within the workflows. For instance, the pipeline waits until all Dimensions are loaded before proceeding, as they are essential for loading the Fact table.

3. **Dataflow ETL (load into DW) of the Dim\_Car table:**



To set up the dataflow containing car information, we imported the "Car and drivers B.csv" file. Since both "Car and drivers A.csv" and "Car and drivers B.csv" contain identical cars, there was no need to load both files. Initially, we promoted the first row to headers and rectified the data types. Since the dataset encompassed information about drivers, we excluded the associated columns. Furthermore, we established a column to function as the car's surrogate key, and although the CarId (BK) column appears sequential, we cannot map the same column into two columns in the DW. Thus, to guarantee a sequential numeric key, we generated the index.

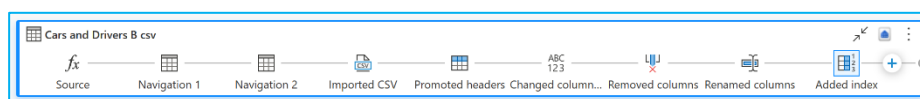
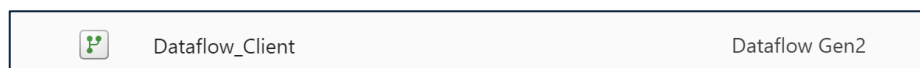


Figure 10. Dataflow\_Car

#### 4. Dataflow ETL (load into DW) of the Dim\_Client table:



When creating the client's dataflow, we began by importing the "customers.csv" file. Through exploration in the notebook, we identified and resolved three issues with this data source. Firstly, we addressed the lack of headers by renaming the columns. Secondly, we standardized the values in the gender column, where 'F' and 'S' represented "Female" clients and 'M' and 'H' represented "Male" clients for clarity. The last issue involved a duplicated client (Client 10000) appearing 10 times. To rectify this, we utilized a group by function on Client ID, Name, Gender, and Birthdate. Finally, similar to the car dimension, we employed an index column as the client's surrogate key.

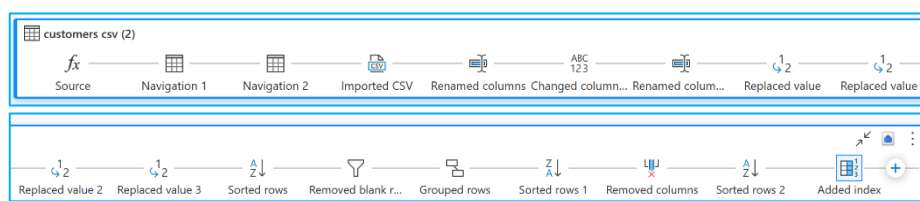


Figure 11. Dataflow\_Client

#### 5. Dataflow ETL (load into DW) of the Dim\_Currency table:



To set up the Currency's Dataflow, we utilized the "NEW ride.csv" as the data source and focused on the "Currency" column. Our objective was to extract the unique values from this column to construct our Currency table. Initially, we observed that only one currency, EUR, was present. However, to accommodate future expansions and payment updates, we introduced conditionals for the Currency name, adding more options for flexibility. This approach ensures that if payments in other currencies are received in the future, their names will be automatically detected without requiring additional steps in the dataflow.

Condition	Operator	Value	Output
If Currency equals EUR	Then	Euro	
Else if Currency equals USD	Then	United States dollar	
Else if Currency equals GBP	Then	Great British Pound	
Add clause			
Else NEW			

Finally, we created a surrogate key using the same method as the dataflows above.

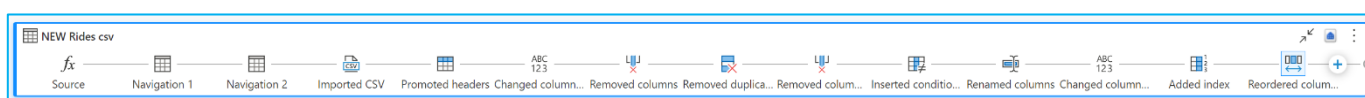
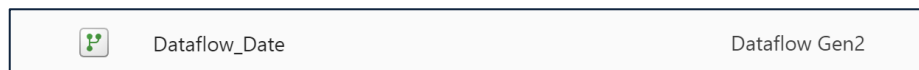


Figure 12. Dataflow\_Currency



*Nova Information Management School, Lisbon, Portugal*

## 6. Dataflow ETL (load into DW) of the Dim\_Date table:



The date dimension is crucial for time-based analysis and reporting within our data infrastructure. To establish a foundation for this column, we began by generating a list spanning from 2020 to the end of 2023:

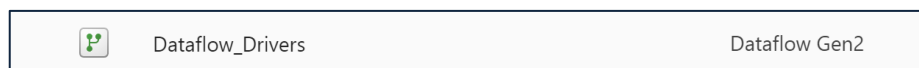
```
List.Dates(#date(2020,1,1), Number.From(#date(2024,1,1)) - Number.From(#date(2020,1,1)), #duration(1,0,0,0))
```

After converting the list into a table, we added essential columns like Year, Month, Day, Week, and Quarter using PowerQuery's 'Date columns' function. Next, we adjusted the day of the week values (Sunday as 7), created semester columns with 'IF' statements, and identified holidays such as Christmas (25/12) and New Year's (01/01) with an 'Is\_Special' column. These steps were aimed at maximizing analysis flexibility. Finally, we completed the process by implementing an intelligent SK (in YYYYMMDD Format) through a merge with a conditional column.



*Figure 13. Dataflow\_Date*

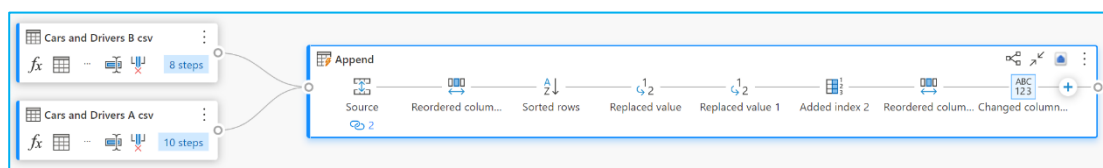
## 7. Dataflow ETL (load into DW) of the Dim\_Drivers table:



We have been provided with two CSV files, Car and Drivers A.csv and Car and Drivers B.csv, which contain information about female and male drivers, respectively.

Initially, we worked on Car and Drivers A.csv. This document contained a lot of quotation marks, which prevented the columns from being automatically assumed. To merge the two sources, we needed them to be in the same format. Therefore, we removed the quotation marks and split the columns by a comma. After that, we merged the two columns into one table, which contained all the information about the drivers.

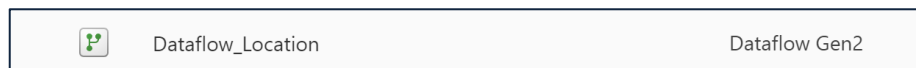
To make the Gender column values easy to interpret and in line with the client's data flow, we changed the values to "Female" and "Male". Finally, we finished the dataflow by creating the sequential surrogate key.



*Figure 14. Dataflow\_Drivers*

Nova Information Management School, Lisbon, Portugal

## 8. Dataflow ETL (load into DW) of the Dim\_Location table:



Although seemingly simple, the Dim Location Dataflow proved to be the most complex among all Dimensions. Initially, we addressed the issue of missing values in the district column by loading the locations of Portugal support file and splitting it into two tables: one containing districts and cities, and the other containing districts and *freguesias* (parishes). This division aimed to facilitate the management of joins. We experimented interactively to ensure optimal results, especially considering that the "cities" column in the Locations file could contain both cities and freguesias.

We first connected to the cities, using **FuzzyNestJoin** with a threshold of 98% to capture cities with a non-case-sensitive search, ignoring accents or trailing spaces in the text. This approach successfully captured the majority of districts. However, we discovered that some cities without districts were actually *freguesias*. Thus, we performed a final outer join with the “Designação FR” table to capture these remaining instances, creating a conditional column to combine the results into a single *District* column. Only one missing value was left and we address this directly, ultimately resolving all missing values.

```
Table.FuzzyNestJoin(("#Locations csv", {"City"}, {"Designação CC"}, {"Designação CC"}, {"Designação CC"}, JoinKind.LeftOuter, [IgnoreCase = true, IgnoreSpace = false, SimilarityColumnName = "Similarity score", Threshold = 0.98])
```

A significant aspect of this Dataflow was addressing the issue of duplicate city with different IDs, such as "Beja," identified earlier. After transforming some incorrect names, we were left with three values—"Beja," "São João da Madeira," and "São Pedro do Sul"—resulting in each city having two IDs.

Value to find	Value to find
<input type="text" value="São João da Madeira"/>	<input type="text" value="Freguesia de São Pedro do Sul"/>
Replace with	Replace with
<input type="text" value="São João da Madeira"/>	<input type="text" value="São Pedro do Sul"/>

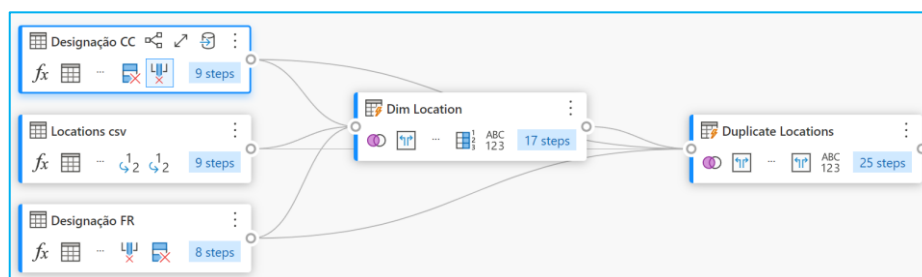
To load the Dim\_Location table, we eliminated duplicates from this column and assigned an index as a sequential surrogate key (SK). However, in the **FK\_Pickup\_Location** and **FK\_Dropoff\_Location** columns of the Fact table, we still had the old IDs<sup>2</sup>. Therefore, to have a way to map this IDs we remove further, we duplicated all transformations up to the step before removing duplicates and created a table solely with duplicates, using the "Keep duplicates" option and the "Is\_Active" column to map the active ID of each city to the "ActiveID" column. Subsequently, we utilized this table to map the values in the Fact Rides Dataflow. We created a support warehouse named **Ride4All\_VALIDATIONSTG\_2024**, making it a two-destination

<sup>2</sup> For Example: Beja had two IDs (18 and 20) in Dim\_Location table. ID 20 was removed to remove duplicates, but references to it in the Fact table column needed to be mapped to ID 18 for data consistency.

*Nova Information Management School, Lisbon, Portugal*

dataflow to load this table. The results of the *Duplicate\_Locations\_to\_map* table can be seen in the Annex.

Despite its apparent complexity, this dynamic approach ensures that even with updates to the city map, all duplicates will be automatically corrected in both the Dimension Location and the Fact Rides.



*Figure 15. Dataflow\_Location*

## 9. Dataflow ETL (load into DW) of the Dim\_Time table:



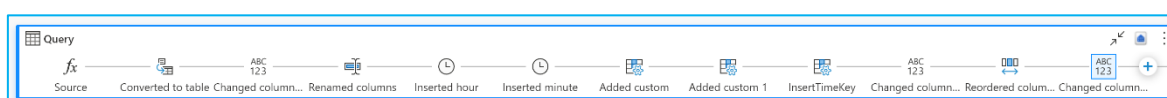
Since we decided to incorporate the flexibility to filter by time of day (as specified in the business questions, for instance, to analyze peak demand hours) into this dataflow, we achieved this by employing a very similar approach to the one used in the Dim\_Date Dataflow. Initially, we compiled a list containing every hour and minute combination of a day:

```
List.Times(#time(0, 0, 0), 24 * 60, #duration(0, 0, 1, 0))
```

After converting this list into a table, we utilized it to add the remaining attributes specified in the schema above, thereby enabling further filtering and display options. However, a challenge arose during this phase as Microsoft Fabric does not support the 'Time' datatype outside of the Power Query editor; it only supports 'DateTime'. We resolved this issue by converting the 'Time' column to a 'Text' format merely to facilitate data storage in the data warehouse (DW).

Subsequently, if the need arises to display this data in visuals that expect numeric values, we can convert this column to 'DECIMAL', where the decimal represents the minutes, or simply utilize the 'Hour' or 'Minute' columns since they are integers. However, for the time being, we have retained it as 'TEXT' to maintain visual consistency.

Additionally, we have implemented an intelligent key for this dimension. Essentially, when the time is, for example, 6:30 PM (18:30), the time key is 1830. In cases where the hour has leading zeros, they are automatically removed since the key is an integer. For instance, 00:00 has the key 0, assuming only the zero on the right.



*Figure 16. Dataflow\_Time*

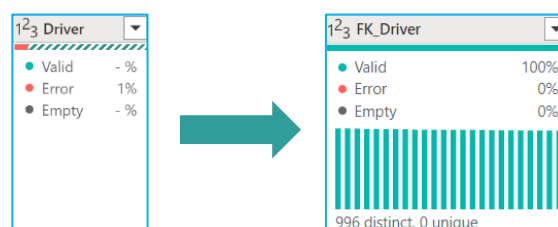
## 10. Dataflow ETL (load into DW) of the the Fact\_Rides table:



Finally, for the last Dataflow, we needed to ensure that all foreign keys in the Fact table had a corresponding match in the dimension tables. To achieve this, we implemented an extended application of outer joins with the already loaded dimensions.

Firstly, we divided the columns containing date-time information (both StartDateTime and EndDateTime) into two separate columns. This allowed each one to have its corresponding foreign key from the Date and Time dimensions, respectively. In the case of the Time table, we applied the technique mentioned earlier, where we transformed the text-format column containing the proper time into a Time column again (inside Power Query).

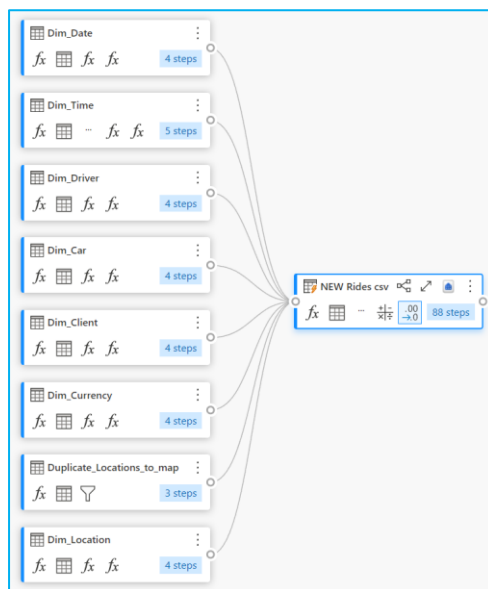
Next, we addressed missing values in the Car column using the Dim\_Driver dimension. Given that each driver has only one car and we have a column specifying the ID of each driver in this dimension, we were able to map the missing value. An important aspect during this phase was having the *Column Profile* view enabled, which allowed us to identify an unidentified issue during the Exploratory Data Analysis (EDA). In the Driver column, we found five IDs with a prefix of "D," causing errors in these cells:



After mapping the Car, Driver, Client, and Currency columns, we moved on to implementing the second phase of our solution for location duplicates. By loading the Duplicate\_Locations\_to\_map column, we transformed the values of non-existent IDs into the active IDs, and then converted them to new values according to the SK of the Dim\_Location table for both the Pickup Location and Dropoff location columns. Once again, this ensured that the mapping was automatic, rather than requiring manual replacement for each value.


Finally, we adjusted data types and rounded our measures to two decimal places (although we may adapt this value in the next phase). We also corrected a value of 7000003 kms, assuming it was an error due to the absence of a comma.

*Nova Information Management School, Lisbon, Portugal*



*Figure 17. Dataflow\_Fact\_Rides*

## 11. Invoke Pipeline Activity:

	PL_Ride4All_VALIDATE_RULES	Data pipeline
---	----------------------------	---------------

***PL\_Ride4All\_VALIDATE\_RULES: Data Quality checks after loading.***

Rule	Description
<i>Validate unique rows in Dimensions</i>	This rule checks if we have a unique combination of attributes across the entire dimension (no rows are duplicated). <sup>3</sup>
<i>Validate unique combination of FK in Fact_Rides</i>	This rule checks if we have a unique combination of Foreign Keys across the entire fact table
<i>Check FK Relationships in Fact_Rides</i>	This rule verifies Foreign Key Relationships between the Fact_Rides table and every Dimension. While we also monitor this during

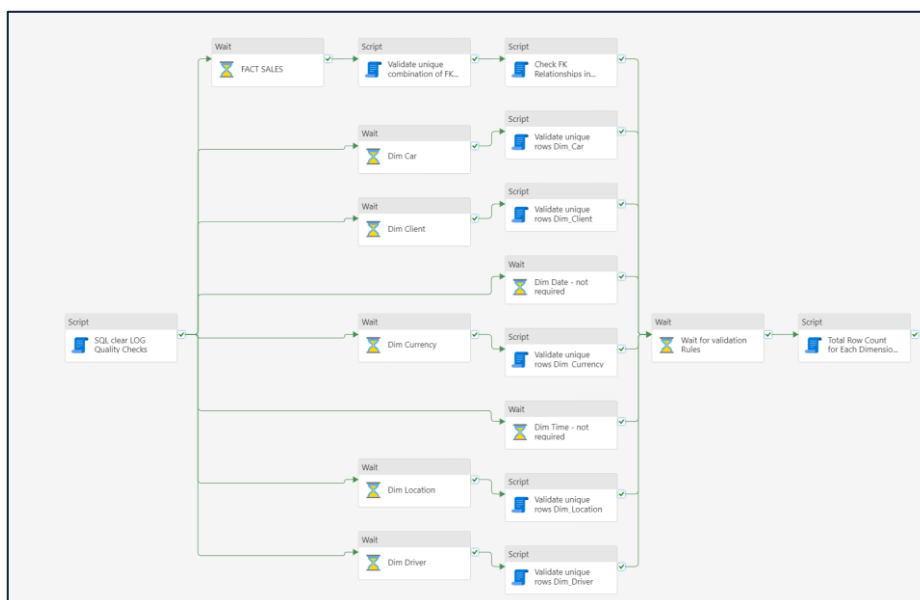
<sup>3</sup> This rule helped us find a duplicate location: São Pedro do Sul. We replaced "Freguesia de São Pedro do Sul" with "São Pedro do Sul" without a space at the end, while another entry "São Pedro do Sul " existed in the table with a space. Since Power Query is sensitive to spaces, it did not recognize them as duplicates, and this rule assumed otherwise. Thus, we removed the trailing space to resolve this issue.

*Nova Information Management School, Lisbon, Portugal*

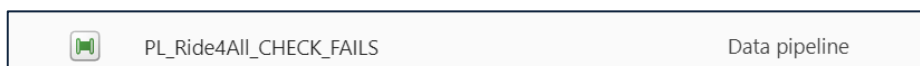
	the outer joins in the dataflow, this rule provides an easy way to double-check it.
<i>Total Row Count for Each Dimension and Fact Table</i>	Technically, this isn't a rule in the traditional sense, as it doesn't produce a result of "FAIL" or "OK." Instead, it automatically retrieves the total row count for each Dimension and Fact Table. This allows us to verify if the counts correspond with the expected values.

*Table 11. Validation Rules*

These results will be loaded into the support DW we created, Ride4All\_VALIDATIONSTG\_2024 in the log\_quality\_checks table, providing an easy way to access them separately from the business data. Additionally, we implemented a second pipeline that iterates over these results.



*Figure 18. PL\_Ride4All\_VALIDATE\_RULES*

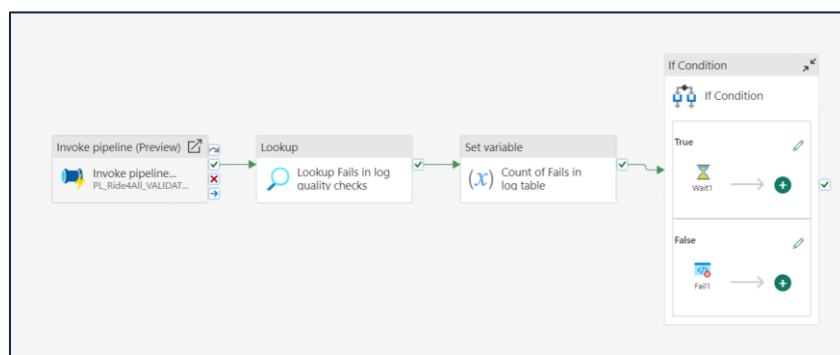


***PL\_Ride4All\_CHECK\_FAILS: Results Checking to conduct the email activity.***

*Nova Information Management School, Lisbon, Portugal*

This extra pipeline iterates over the validation results and counts the number of FAILS (if any) in the 'etl\_result' column of the **log\_quality\_checks** table using a **Lookup** activity. The results are stored in a variable. If there are FAILS, meaning the count is greater than zero, the pipeline is marked as unsuccessful. In this case, we receive an email prompting us to investigate which rule failed and in which dimension, allowing us to correct the ETL process. If everything is fine, we receive an email confirming that the data warehouse was loaded successfully and that all validation rules passed.

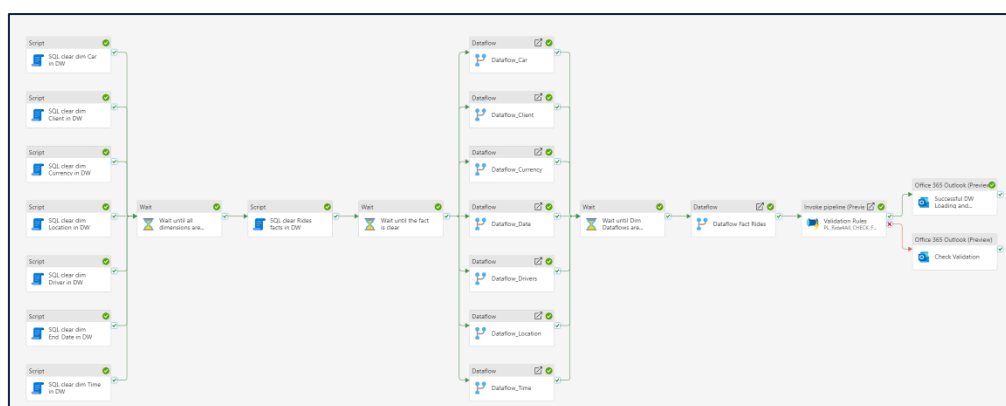
Including this functionality facilitates monitoring the pipeline's status, particularly when scheduling loadings. This proactive measure enables swift reaction and issue resolution in the event of any errors during the loading process.



*Figure 19. PL\_Ride4All\_CHECK\_FAILS*

## 12. Final Design:

As depicted in the final design, it's crucial to ensure that all dimension tables are loaded before initiating the Dataflow Fact Rides. Otherwise, the process will not function correctly, as the Fact Rides Dataflow relies on the dimension tables being already loaded. This is why the Wait activity is so useful at this stage—it allows us to pause the pipeline until the necessary dimension tables are fully loaded, ensuring smooth and accurate execution of the subsequent steps.



*Figure 20. PL\_Ride4All\_LOAD\_DW*

*Nova Information Management School, Lisbon, Portugal*

## 6 Semantic Model

### 6.1 Connection to Power BI Desktop

When connecting our report to our data, we had several options available. However, we needed to make a decision that aligned with our needs at this stage.

#### 1. Power BI Direct Query Connection

With Direct Query, the semantic model directly queries the data source at runtime. No data is stored in Power BI, so we always query the data present in the data source itself, meaning there is no Data View option available. However, given that we wanted to have complete control over the project and access to the data while working on the report to remind ourselves of the possibilities, we found this mode limiting. Additionally, Direct Query has significant limitations regarding the use of *Time Intelligence*. At this stage, we decided that we wanted more flexibility in terms of functionalities.

#### 2. Import Mode

In this option, we would connect to the semantic model built in Microsoft Fabric, but this time we would select Import. Although this option was initially chosen for offering us a broader range of possibilities when building the report, we ultimately decided to take a different path.

#### 3. SQL Server Connection<sup>4</sup>

This was the chosen option as it offered us the greatest flexibility. We decided to build the semantic model in Power BI (even though we had already started a model in Microsoft Fabric) for two main reasons. The first reason was purely preventive: we encountered some initial bugs when starting the construction in Fabric, and to prevent issues at a later stage due to capacity problems, building in Power BI provided us with more security in this aspect. The second reason was that we wanted to have complete control to make changes throughout the report construction. Additionally, we wanted all the analytical capabilities that Power BI offers (creating groups, clusters, new tables, etc.), and adjustments to the semantic model would be instantly effective without needing to switch tools and refresh each time, such as for sorting columns. This improved our workflow and allowed us to focus more on optimizing the model at this stage.

It's worth noting that once the model is optimized, we can always move it to Fabric and establish connections differently. For us, being in a learning and experimentation phase, this method made the most sense.

---

<sup>4</sup> [75u33zhx4yxezmshig2uxisjby-4hs6nzwni3sefkbwkolhatjrgm.datawarehouse.fabric.microsoft.com](https://75u33zhx4yxezmshig2uxisjby-4hs6nzwni3sefkbwkolhatjrgm.datawarehouse.fabric.microsoft.com) (thi is our DW link)



*Nova Information Management School, Lisbon, Portugal*

## 6.2 Model optimization

Needing to build our model from scratch, we made sure that:

- The relationship is one (in the dimension) to many (in the fact table).
- We assume referential integrity in relationships.
- We connect D Date and D Time with two relationships (Start Date and Start Time active, and End Date and End Time inactive). The Location Dimension will be approached differently.

Then, to simplify the model's appearance, we renamed the dimensions and the fact table (adding a prefix D or F) as well as the columns (attributes of each table). The important thing here was to ensure that the names were simple and intuitive to understand. While doing this, we hid the technical fields/columns by clicking "hide" so they wouldn't appear in the report view (considering that most of them won't be useful for the report). This way, we simplified our view of the available fields and made our work more organized.

- **D Date Table:**

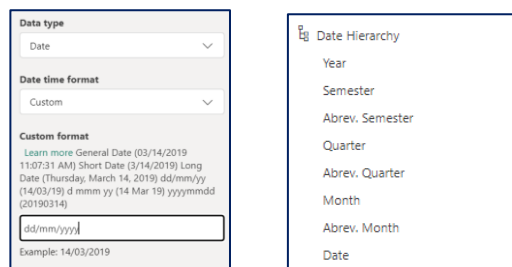
✓ Mark as date table

One important step at this stage is to mark our Date Dimension (D Date) as the date table. We then marked the Date column with a custom date format (dd/mm/yyyy) and sorted the text columns according to the corresponding numerical columns. For example, it is important to sort the *Month* and *Weekday* columns according to the *Month Number* and *Weekday Number* columns to prevent them from being sorted alphabetically in the visualization. This way, we can analyze temporal trends sequentially.

We defined the **Date hierarchy** according to what was specified in the hierarchies section. However, here we ended up adding both the full names and abbreviations for *Semester*, *Quarter*, and *Month* to manage whether we want abbreviated names or not when selecting the hierarchy and its elements for visualization. (In the end, we hid the columns that belong to the hierarchy because we don't need to have these fields separately).

During the report construction, we noticed that it would be interesting to have a forecast by month over the years. Since we couldn't use the forecast function with the hierarchy, we created a **calculated column** called **Month Year Combo** to provide the time horizon for aggregating rides by month and performing the forecast.

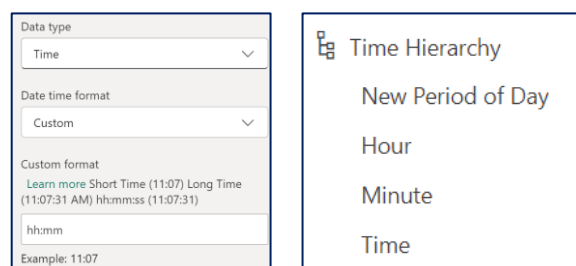
*Nova Information Management School, Lisbon, Portugal*



*Figure 21. D Date Optimizations*

- **D Time Table**

Recalling a limitation faced during the data transfer related to the time of each ride, where we couldn't define our *Time* column as Data Type *Time*, now during the model construction, this limitation can be overcome. We switched from the TEXT format to a custom Time format (hh:mm). An important step here was using the *Hour Range* column to define a new calculated column called *New Period of Day*, as the previous *Period of Day* column had hours disproportionately distributed, affecting the comparison of periods of the day. In the new column, we ensured 6 hours in each group (Morning, Afternoon, Evening, and Night)<sup>5</sup>.



*Figure 22. D Time Optimizations*

- **D Dropoff Location and D Pickup Location Tables**

The decision to create a calculated column from the original D Location will become more justified in the technical explanation part of the report. However, this step was adopted because we need two different sources of location data (Origin and Destination) to create the Flow Map. Thus, each of these dimensions is linked to its

<sup>5</sup> The decision not to alter the Period of Day column directly in the Dataflow that loads the information into the Data Warehouse was made because, although disproportionate, this column was defined according to the standard division of the periods of the day. The new column, New Period of Day, with equal interval periods, was a visualization decision. Therefore, it was altered only in Power BI, providing the flexibility to use different divisions according to the analytical preference of the report author.

corresponding FK in the fact table, allowing us to perform the desired analysis. We highlight that the locations were converted to the Data Category Country or City, and the Location Hierarchy was created in each of these dimensions (Pickup Location Hierarchy and Dropoff Location Hierarchy).

D Pickup Location	...	D Dropoff Location	...
BK_Location		BK_Location	
Country		Country	
District		Dropoff City	
Pickup City		Dropoff District	
SK_Location		SK_Dropoff_Location	
Pickup Location Hierarchy		Dropoff Location Hierarchy	
Pickup Country		Dropoff Country	
Pickup District		Dropoff District	
Pickup City		Dropoff City	
Collapse ^		Collapse ^	

Figure 23. D Dropoff Location and D Pickup Location Optimizations

- D Car Table**

In this dimension, we only hid the technical fields and defined the Car Hierarchy to be used in our operational analysis.

Car Hierarchy
Manufacturer
Brand
Year of the Car
Plate Number

Figure 24. D Car Optimizations

In the D Client and D Driver tables, the transformations were quite straightforward. We hid the technical fields and the birthday dates since we won't use them in the analysis (we could, for example, use them to define clusters, but in this project, we will take a simpler approach and define the clusters according to the measures). The D Currency table was totally hidden since we don't need it at this stage.

*Nova Information Management School, Lisbon, Portugal*

### 6.3 Measures and calculated columns

In addition to the calculated columns mentioned above in each dimension (which remain as attribute columns), we now move on to the measures and calculated columns that we added to our fact table, F Rides. In the following tables, you can find the description of each measure and calculated column:

Name	Description
<b>Avg Amount Per Trip</b>	Gives us the average amount per ride. (KPI)
<b>Avg Duration of Rides (min)</b>	Gives us the average duration of the rides in minutes. (KPI)
<b>Avg Kms Per Trip</b>	Gives us the average distance of the ride in kilometers. (KPI)
<b>Avg Rides Weekday</b>	Gives the average number of rides on weekday. (Time Intelligence)
<b>Avg Rides Weekend</b>	Gives the average number of rides on weekends. (Time Intelligence)
<b>Avg Number Of Rides</b>	Gives us the average number of rides of the drivers. (KPI)
<b>Avg Number Of Rides All Drivers</b>	Gives us the average number of rides made by all the drivers, ignoring selections. (KPI)
<b>Car First Ride Date</b>	Gives the date of the first ride of each car. (Time Intelligence)
<b>Car Utilization Rate</b>	Gives the the proportion of days in which the car was used in relation to the total number of days in the period under analysis. (Time Intelligence)
<b>Client First Ride Date</b>	Gives the date of the first ride of each client. (Time Intelligence)
<b>Client Last Ride Date</b>	Gives the date of the last ride of each client YTD. (Time Intelligence)
<b>Driver First Ride Date</b>	Gives the date of the first ride of each driver. (Time Intelligence)
<b>Driver Last Ride Date</b>	Gives the date of the last ride of each driver YTD. (Time Intelligence)
<b>Max Number of Rides All Drivers</b>	Gives the number of rides made by the driver who did more rides, ignoring selections. (KPI)

*Nova Information Management School, Lisbon, Portugal*

<b>MoM% Number of Rides</b>	Give the percentage variation in the number of rides per month. (Time Intelligence)
<b>Number of Clients</b>	Shows the total number of clients who took rides in a specific context. (KPI)
<b>Number of Drivers</b>	Shows the total number of drivers who took rides in a specific context. (KPI)
<b>Number of Rides</b>	Gives the total number of rides. (KPI)
<b>Number of Rides All Locations</b>	Gives the total number of rides of all locations, ignoring selections. (KPI)
<b>Total Amount</b>	Gives the total amount. (KPI)
<b>Total Kms</b>	Gives the total kilometers. (KPI)
<b>YoY% Rides Variation Dinamic</b>	Percentage change in the number of rides compared to the same period in the previous year, calculated dynamically. (Time Intelligence)
<b>YTD Amount Current Year</b>	Year-to-date total amount for the current year. (Time Intelligence)
<b>YTD Amount Last Year</b>	Year-to-date total amount for the previous year. (Time Intelligence)
<b>YTD Rides Current Year</b>	Year-to-date total number of rides for the current year. (Time Intelligence)
<b>YTD Rides Last Year</b>	Year-to-date total number of rides for the previous year. (Time Intelligence)

*Table 12. Measures*

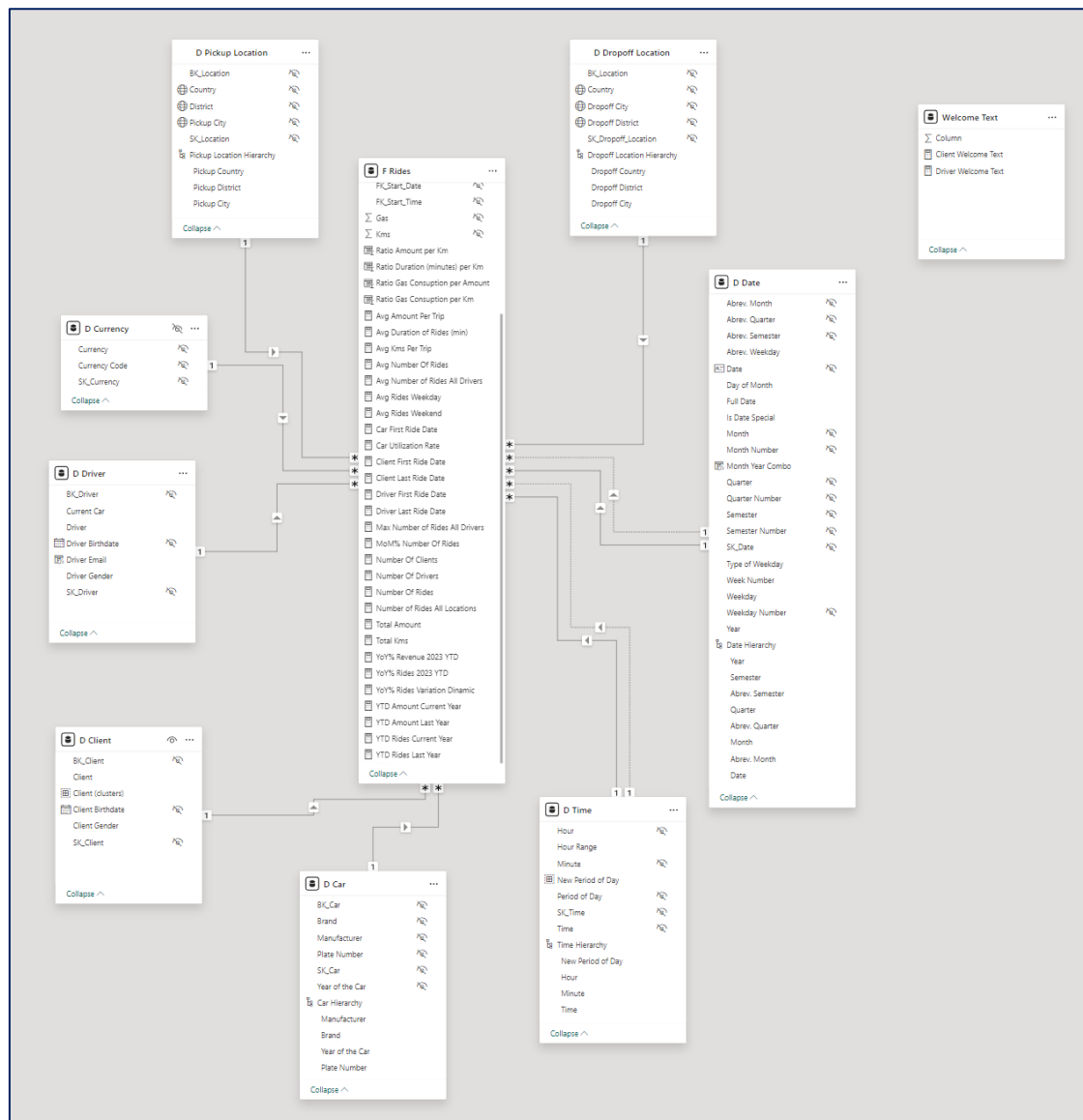
Name	Description
<b>Ratio Amount per Km</b>	Gives the average amount per kilometer per trip row.
<b>Ratio Duration (minutes) per Km</b>	Gives the average duration per kilometer per trip row.
<b>Ratio Gas consumption per Amount</b>	Gives the average gas consumption per amount per trip row.
<b>Ratio Gas consumption per Km</b>	Gives the average gas consumption per kilometer per trip row.

*Table 13. Calculated columns*

*Nova Information Management School, Lisbon, Portugal*

As an additional note, we created a new column called Welcome Text solely to store some DAX formulas that customize the welcome message for the Driver Profile and Customer Profile. Although we adopted the approach of storing all our measures in the Fact Table, since these are purely aesthetic, we chose not to add them along with the others to keep them contextually appropriate.

**Final Result:**



*Figure 25. Final Semantic Model*

## 7 Power BI report<sup>6</sup>

Taking into account the business needs and questions mentioned earlier, a decision has been made to divide the report into sections so that each of these sections has its own narrative in order to maintain the principle of proximity of the insights and help maintain a cohesive storytelling. On the cover of the report, we can find this division:

- **2023 Overview:** This Dashboard represents a summary of all the following reports. It presents information related only to 2023 YTD and has the main objective of providing a quick and intuitive view of the company's current status.
- **Performance Analysis and Forecast:** Here we have a more comprehensive performance analysis temporally covering historical data and having greater flexibility to perform analyses with different time horizons. Additionally, we can find the forecast based on trends in historical data.
- **Driver Analysis (Driver Profile):** This section aims to create a way to analyze the performance of each driver.
- **Client Analysis (Client Profile):** Following the same reasoning and addressing one of our client's requirements, we have created in this section a possibility to analyze and create a profile for each client.
- **Location Analysis:** An essential part of our project was to provide Ride4all the flexibility and tools necessary to thoroughly analyze and optimize their routes. Therefore, we have a section to explore this possibility.
- **Operational Analysis & Car Analysis:** This section introduces the possibility of analyzing the physical resources of the company in order to identify possible operational improvements.

Furthermore, recalling the needs highlighted by the client, we have the following essential features in our project:

- **Trips per month, Quarter, and Yearly (accumulated to date):** Can be found in the "2023 Overview" section (only for 2023), "Performance Analysis," and can also be found and filtered by driver in the "Driver Analysis" section.
- **Number of Rides as well as Average Distance of Rides:** These key performance indicators (KPIs) are available in the "2023 Overview", "Performance Analysis", and across the report, with the option to filter based on different contexts within the sections.
- **Comparison of current year vs previous year:** This analysis is enabled throughout the report, especially in the "Performance Analysis" section.

---

<sup>6</sup> One important note is that we filter the entire report to show only the data up to May 2023 since it's the last complete month we have. June only has a few days of data, which can be misleading in terms of visualization.

*Nova Information Management School, Lisbon, Portugal*

Additionally, except for the “2023 Overview” dashboard, the remaining sections allow filtering by years, allowing for comparison within different contexts.

- **Building client profiles:** This process is straightforward with the creation of the “Client Analysis (Client Profile)” section.
- **Ride details (at the trip level):** In the “Client Analysis (Client Profile)” section, the information begins with the client, and through interacting with the relevant visualizations, we can delve into the details of the information for each specific trip.
- **Aggregate Rides by driver, pick-up point, and destination point:** The “Driver Analysis (driver profile)” section provides these possibilities. Start by selecting the driver, then choose the district or city of pickup on the map, and finally analyze the respective routes taken by the driver.

## 7.1 2023 Overview

### Main technical aspects

First, considering the potential audience for this section, the aim was to develop a dashboard containing a summary of the information from the rest of the report, but focusing only on the current year (05/2023, excluding the first days of June) to serve as an executive summary, for example, for the CEO. In this case, it would serve, for instance, to keep the CEO updated monthly on the overall company situation in an intuitive and direct manner, considering that these personas typically have limited time and need quick updates to make rapid decisions.

Therefore, in the top section, we can find a line with the essential KPIs of the analysis (Number of rides, year-over-year growth in rides and earnings YTD, Average distance of rides, among others). After that, we move on to a Trend Comparison using a line chart to track the evolution of rides and earnings across the year, a column chart to analyze the distribution of rides during the days of the week (in %), and a line chart to evaluate the evolution of ride demand throughout the hours of the day. Due to the number of drivers and locations, we decided to use a subsection to show the TOP 5 in a more intuitive way to identify top performers and areas of higher demand in terms of pickup and drop-off.

### Key findings

In this section, besides the obvious **YTD growth** in 2023 in both the number of rides (+19.59%) and earnings (+19.81%), we can see that we have approximately **167K rides** YTD in 2023 and **1.71M euros** in earnings YTD 2023. Additionally, we were able to address the following questions from our stakeholders.



Nova Information Management School, Lisbon, Portugal

1. **Who are the top 5 drivers in terms of total rides in 2023?** Francisco Eduardo Moreira (380), Gabriel Fábio Rodrigues (376), Bruno António Almeida (373), Carlos Bruno Pereira (371), and Nuno Martim Ribeiro (371).
2. **What is the total number of drivers who have completed rides Year-to-Date (YTD) in 2023?** 980 Drivers (by drilling up, we can see the total drivers who completed rides in 2023 in the tooltip of the line chart).
3. **What are the top 5 performing locations in terms of the number of rides in 2023?** In terms of pickup districts, we have Aveiro, Porto, Viseu, Braga, and Coimbra. Similarly, in terms of dropoff districts, we have Aveiro, Porto, Viseu, Braga, and, differing, Faro closing the top.
4. **What are the most demanded hour ranges of the day based on the number of rides and the start time of the ride, in 2023?** We can clearly see the growing demand between 7 am and 9 pm, with two peak demand points at 8 am and 8 pm.
5. **What is the average fare per trip for the top driver in 2023?** €10.98 for Francisco Eduardo Moreira.
6. **What is the peak hour (in terms of the number of rides) at the top pick-up District (in terms of the number of rides) during the first quarter of 2023?** 8 am.

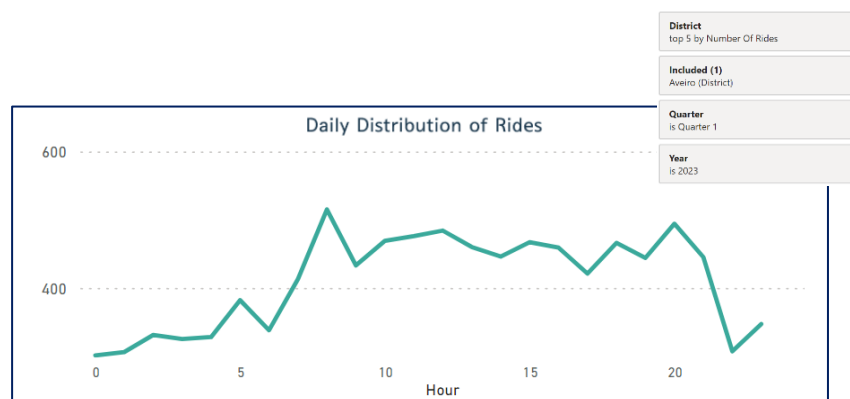


Figure 26. Daily Distribution of Rides (Aveiro, Q1 2023)

## 7.2 Performance Analysis and Forecast

### Main technical aspects

As mentioned before, this section of the report is divided into two parts. First, we have the performance analysis, a general analysis that can be filtered across the historical years of the data. The second part contains the forecast for the rest of 2023 and the beginning of 2024. On the first page, we find the KPIs and the TOPs in the upper section, including the top pickup and drop-off locations, drivers, top month, day of the week, and hour range. Additionally, we have the average number of rides on weekdays and weekends. Following that, there is a line chart to compare the evolution of rides over time, and a matrix to cross-reference the days of the week and times of the day (the cells are color-coded gradually, with periods of higher ride demand being darker

*Nova Information Management School, Lisbon, Portugal*

and vice versa). Finally, we have two KPI charts to analyze the evolution of rides and earnings between the current period and the previous period (YTD).

On the forecast page, we present annual, daily, monthly, and monthly variation forecasts, though the most comprehensive is the monthly one, which shows a clearly demarcated and easy-to-read variation. We included the other forecasts to offer flexibility.

### **Key findings**

**Excluding 2023**, we observe a consistent growing trend across the years. Additionally, August is the busiest month, Saturday is the busiest day of the week, and 8 – 9 PM is the peak time during the day. The average number of rides on weekends and weekdays remains fairly consistent over time.

1. *Who was the best driver in terms of total rides in 2022?* Martim Luís Pereira (746 rides).
2. *What is the most demanded hour of the day on the busiest day of the week based on the number of rides and the start time of the ride across the years (2020-2022)?* It is consistently 8 – 9 PM.
3. *What is the average number of rides percentage variation between weekdays and weekends in the first quarter of 2023?* 14.33K rides on weekdays and 13.77K rides on weekends:

$$\text{Percentage variation} = \left( \frac{14.33K}{13.77K} - 1 \right) \times 100$$

There are about 4% more rides on weekdays than on weekends in the first quarter of 2023.

4. *What is the average duration of rides in the busiest month of 2022?* The busiest month in 2022 was July, and the average duration of rides was 51.51 minutes.
5. *What is the total projection (forecast for the number of rides) for 2023?* 376,239 rides.

## **7.3 Driver Analysis (Driver Profile)**

### **Main technical aspects**

This subsection of the report is specifically designed to provide a performance analysis tool for each driver. That said, we have the KPIs related to each driver at the top of the section, including two cards showing the date of the driver's first ride and the date of their last ride (YTD). Additionally, we provide the distribution of the number of rides over time, the distribution of rides throughout the week (with a tooltip that shows the daily distribution over each day), and a map displaying the driver's pickup locations (where larger and darker bubbles indicate more rides). An important feature is that by clicking on one of the points on the map, the "Go to Driver's Routes Analysis" button becomes enabled, allowing us to navigate to a page where we can analyze the routes for each specific ride, thus fulfilling the need to aggregate rides by driver, pick-up point, and destination point.

*Nova Information Management School, Lisbon, Portugal*

## Key findings

Analyzing the Best Driver of 2022 in terms of the number of rides, Martim Luís Pereira:

1. ***What is the total number of rides (all years)?*** 2,188 total rides.
2. ***What is the period of the day with the most rides on the day of the week with the highest percentage of rides (all years)?*** Thursday has the highest percentage, accounting for 14.99% of the weekly rides for this driver. The tooltip shows that the afternoon is the busiest time of the day.
3. ***What is the average ride distance in kilometers in the last month of analysis (May 2023)?*** 48.70 km.
4. ***In the pick-up district with the most rides (all years), what is the most repeated route?*** The pick-up district with the most rides is Aveiro with 240 rides, featuring three routes with 5 trips each: Anadia -> Anadia, Ílhavo -> Ílhavo, and Pardilhó -> Pardilhó.

A note for this section: *Row Level Security (RLS)* was developed for each driver, and this process is detailed later in the report.

## 7.4 Client Analysis (Client Profile)

### Main technical aspects

Fulfilling the need to create a client profile, we combined utility with convenience in this page, allowing for the analysis of ride details at the trip level. Since a ride only begins with a client's request, by selecting a client profile, we can analyze the details of the trips they have taken.

This includes their average spending, typical ride duration, usual travel distance, total number of rides, and the first/last ride date to determine if they are a new or a possible churned client. We also created clusters to divide the clients according to the number of rides taken. Additionally, we can identify their most frequent drivers and create a schedule by day of the week and hour range to understand the peaks of demand for each client. Furthermore, we can track the route of the client's trips from pick-up to drop-off to understand popular routes.

Finally, a KPI compares the client's rides this year to last year to easily detect any potential churn alerts, thereby motivating possible promotional efforts to encourage the customer to use our service more frequently.

*Nova Information Management School, Lisbon, Portugal*

### Key findings

Considering the profile of **Client 10**, who has a total of 115 rides, we observe a substantial growth YTD with 17 rides in 2023 compared to 6 rides in the same period in 2022 (+183%):

1. ***Who are the drivers with the most repeated rides from this customer (all years)?***  
 There are two drivers with the most repeated rides.
2. ***What is the route and the driver of the most expensive trip taken by the customer?***  
 The trip with the highest fare was from Peniche to Mira with driver Luís Leonardo Moreira, costing €42.39.
3. ***On the day of the week with the most rides by the client, what is the hour with the most rides and what is the average ride duration of those rides (all years)?***  
 The day with the most rides is Saturday with 23 rides. The busiest hour is 1 – 2 PM with 3 rides, and these rides have an average duration of 50.67 minutes.

Other applications can include identifying if a client frequently repeats the same route to create specific promotions for that route to encourage them to choose Ride4All. For Client 10, there are no repeated routes.

## 7.5 Location Analysis

### Main technical aspects

The Location Analysis was constructed to help Ride4All gain insight into how their business is performing across various areas, enabling data-based analysis and strategic decision-making.

To start, we have two filters where it's possible to choose the district where the ride started and ended. After making a selection, the report updates all the information for those specific locations.

At the top of the report, the driver and the three cars with the highest number of rides in the selected locations are displayed, along with key data about those rides: average ride amount, average ride duration, and average ride distance. This report also shows information about the number of rides between the chosen locations and their proportion of the total rides. It also displays the evolution of the number of rides over the years and a projection for the next 12 months.

Additionally, there is a matrix crossing the day of the week and the period of the day to better understand the trends of rides between these locations throughout the week (darker shades indicate higher ride volumes, and lighter shades indicate fewer rides). It also presents the specific pickup and drop-off cities and a map that shows the rides.

Nova Information Management School, Lisbon, Portugal

## Key finds

1. **What is the period of the day with the most rides in the top pick-up location?** Knowing that the top pick-up location is Aveiro, the period of the day with the most rides is the afternoon. This coincides with the time of day when there are more rides in general.
2. **Considering the top pick-up location (district), what is the top drop-off location (district) with the highest number of rides and which day of the week has the highest demand for this route?** As previously mentioned, the top pick-up location is Aveiro. Coincidentally, the drop-off location with the highest number of rides is also Aveiro, indicating a large volume of intra-city travel within Aveiro. The day of the week with the most rides for this combination is Monday, with 2,920 rides.
3. **Within the combination from the previous question, what is the route (by cities) with the highest number of rides and what is its average fare per trip?** The route with the highest number of rides is São João da Madeira -> São João da Madeira (1,900 rides) with an average fare per trip of €10.02.
4. **What is the forecast in terms of the number of rides for the next full month (06/2023) for the route mentioned in the previous question?** The forecast is 59 rides in June 2023 for the São João da Madeira -> São João da Madeira route.

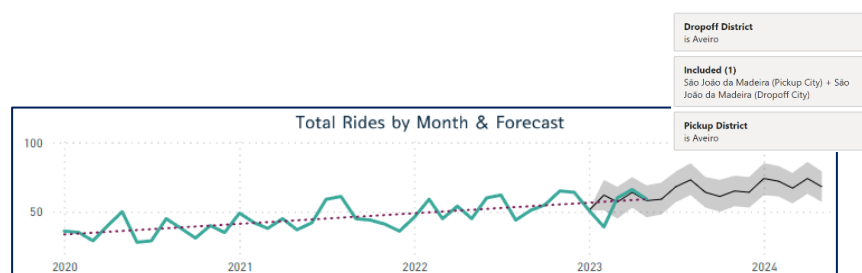


Figure 27. Rides forecast (June 2023 for the São João da Madeira -> São João da Madeira route).

## 7.6 Operational & Car Analysis

### Main technical aspects

The Operational Analysis section provides valuable insights into the performance of our business at an operational level.

First, we present a donut graph where each slice represents a manufacturer. Each slice includes the car hierarchy (manufacturer, brand, year of the car), and its size represents the proportion of rides that cars in each category have completed out of the total rides. This is valuable for Ride4All to understand the company's physical portfolio of cars. This report presents the total number of rides, kilometers traveled, and earnings, as well as three key ratios: average duration per kilometer, average gas consumption per

*Nova Information Management School, Lisbon, Portugal*

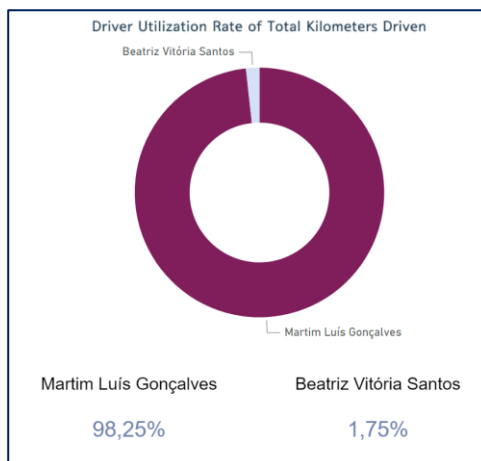
kilometer, and average gas amount. These ratios offer insights into operational efficiency, ride profitability, and vehicle performance.

On the right side of the dashboard, we can see all the cars listed by their plates and the number of rides each has completed, organized in descending order. By selecting one of these cars and pressing “Go to Car Analysis,” the view changes to a dashboard presenting detailed information about the chosen car, facilitating deeper analysis and understanding of a specific car. The Car Analysis dashboard shows specific information about the car, as well as totals and ratios similar to the previous dashboard but associated with this particular vehicle. It also includes a usage rate to see how many days the car has been operational (with rides) within the defined time frame. Additionally, a central feature of this section is the ability to analyze the driver utilization rate of total kilometers driven by male and female drivers to determine if one group typically drives more kilometers. This section also includes the weekly distribution of rides and the top pick-up and drop-off locations frequented by the car during its rides, providing a comprehensive analysis of each vehicle in the Ride4All fleet.

### **Key Findings**

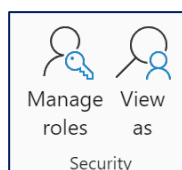
1. ***What are the top 5 manufacturers in terms of number of rides?*** Seat, Peugeot, Opel, Nissan, and VW.
2. ***What is the average gas consumption per kilometer and the average gas consumption per amount for the manufacturer with the most rides?*** The average gas consumption per kilometer is 0.07l/km, and the average gas consumption per amount is 0.52l/€. Note that these ratios are quite consistent across the dataset, which may not make sense in the scope of the analysis.
3. ***What is the manufacturer with the highest average gas consumption per km?*** All the manufacturers have the same average gas consumption per km (0.07l).
4. ***What are the top 3 cars with the highest number of rides?*** AZ-38-ZY, an Audi A4 of 2017, driven by Martim Luis Gonçalves and Beatriz Vitória Santos; HG-38-GF, a Mercedes C220 of 2020, driven by Alexandre Vasco Jesus and Natália Mariana Rodrigues; VU-50-UT, a Nissan Leaf of 2018, driven by Joaquim Hugo Cardoso and Joana Filipa Rodrigues.
5. ***What are the top 3 cars with the lowest number of rides?*** YX-62-XM, an Opel Crossland of 2017, driven by Gabriel Fábio Machado and Beatriz Vitória Costa; VU-78-UT, a Seat Ibiza of 2019, driven by Bernardo André Ferreira and Joana Filipa Cardoso; SR-87-RQ, an Opel Crossland of 2018, driven by Tiago Rodrigo Ribeiro and Sara Natália Silva.
6. ***In the car with the most rides, what is the usage rate of kilometers for each driver?*** For the Audi AZ-38-ZY, driver Martim Luis Gonçalves has a utilization rate of 98.25% of the kilometers driven, while Beatriz Vitória Santos has only 1.75%, showing a significant discrepancy in performance.

*Nova Information Management School, Lisbon, Portugal*



*Figure 28. Driver Utilization Rate of Total Kilometers Driven*

## 8 Extra: Row Level Security



One of the extra requirements our client referred to was the need for a solution that implements Row-Level Security of Rides by User/Location.

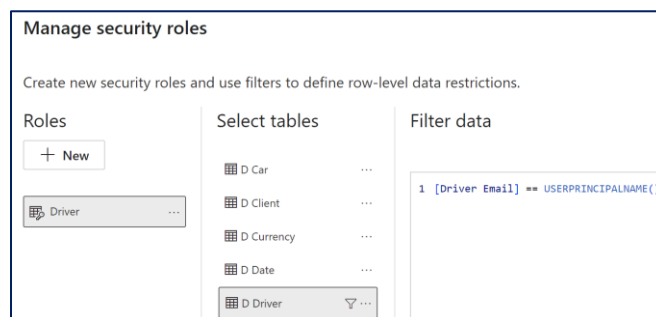
In this project, we implement **row-level security for each Driver** to allow this report to be shared with all drivers without the risk of leaking information intended only for higher levels or sharing the performance of other drivers (such as seeing that driver X is very successful in location Y and starting to steer their work towards that location). Thus, each driver will have a tool that allows them to analyze their individual performance without any conflict of disclosed information.

To proceed with this implementation, we decided to follow industry standards and use each driver's personal email to control access. Since this is not included in the initial data of the Drivers' information, we created a calculated column just to demonstrate the applicability of this procedure in a real scenario where we could request employees' emails. Basically, we use the drivers' names, put everything in lowercase, separate the names with a ".", and end with @ride4all.pt, which would be what would happen in a real scenario.

Driver Email = LOWER(SUBSTITUTE('D Driver'[Driver], " ", ".")) & "@ride4all.pt")

*Nova Information Management School, Lisbon, Portugal*

In the option **Manage roles** in the modelling view, we introduce the following DAX filter:



*Figure 29. Row Level Security filter for Drivers*

Using `USERPRINCIPALNAME()`, we can then select **View As** and introduce an email of an individual driver. The result is that all the info is then filtered for that specific driver, and he only sees his own information. This way, Ride4all gains an extra layer of security for its data.



*Figure 30. Row Level Security Example*

In the future, regarding Row-Level Security by location, other alternative uses could involve the same reasoning but for different countries. For example, if Ride4all expands to Spain, the managers in Portugal would only see information related to Portugal. Or, if we had information about the territory of each driver (e.g., a driver allocated to the northern part of the country), we could do the same. However, we currently do not have information related to this, and all drivers seem to have quite dispersed activity throughout the country, so it didn't make sense to go into that level of filtering for now. If the company provides more information, it will be easily adaptable.



*Nova Information Management School, Lisbon, Portugal*

## **9 Learning Curve & Critical Review**

### **9.1 Date, Time & Location Dimensions**

During the ETL phase, we decided to alter our initial approach. As we continued working with Fabric and researching, it became evident that adopting only one dimension for Date, Time, and Location was more beneficial, especially when building dataflows and loading data. Recognizing the transformation possibilities and the flexibility to manipulate data that Fabric provides, we implemented the solution as specified on pages 16 and 17. This approach not only helps manage workload more effectively but also improves the runtime of pipelines, storage utilization, and optimization by reducing data duplication.

However, during the construction phase of the report in Power BI, we encountered some challenges. While it is possible to manage the Date and Time dimensions with just one dimension connected to the fact table with an active and an inactive relationship, the Location dimension required a different approach in this final phase.

One of our visualization goals was to use a Flow map to analyze different routes throughout the report, which requires location inputs from different sources (one for the Origin and another for the Destination). We took advantage of our SQL Server connection in the project (as explained earlier) and created a new table based on the Location dimension. This allowed us to use one table for D Pickup Location and another for D Dropoff Location.

This approach enabled us to achieve the desired result. However, the decision-making process was in constant flux throughout the project, evolving with the needs and requirements of each phase and with the insights we gained over time.

### **9.2 Period of the Day Definition**

The initial definition of the period of the day was incorrect (for example, the Night period had more hours than the other periods, leading to the misleading impression that most trips occurred at night because we were not considering equal intervals of 6 hours each). We had to use the Groups function to define a new column. It is important that these intervals have the same range for accurate comparison. When we reach the final phase of presenting and utilizing the data, it is easier to have a more holistic and fundamentally unified view of the data and our model. Therefore, it is often necessary to go back and forth to achieve the most efficient result.

*Nova Information Management School, Lisbon, Portugal*

### **9.3 Map Visualizations**

Another important note is that, although our locations are categorized correctly (City & Country), some visualization issues can arise. For example, in the standard Map visualization in Power BI, some Portuguese cities were incorrectly categorized as being in Brazil and were excluded when viewing the distribution by city. Additionally, when using Flow Map, sometimes the routes do not appear automatically (this could be due to the number of rides on the route being insignificant in the total dataset). To address this, all Flow Maps include a supporting table, and by clicking on the specific trip we want to see, it forces the map to display the desired visualization.

### **9.4 Constant Ratios**

One particularity of the dataset, which we had already noticed during the exploration phase, is that some ratios remain constant throughout all trips. For instance, the gas consumption per kilometer is typically 0.07l/km with minor discrepancies. The same occurs with other ratios that show little variation across the dataset (trips from Porto to Lisbon or from Aveiro to Porto have a similar average number of kilometers per trip). Although this is recognized as a particularity of the dataset itself, we note that in a real-world scenario, these ratios could be important drivers for more specific analyses. For example, if a car is consuming too much gas per kilometer, it might be beneficial to replace the vehicle for greater efficiency.

*Nova Information Management School, Lisbon, Portugal*

## **10 Conclusion: Main aspects & Key takeaways**

In this project, we began by addressing Ride4ALL's business challenges and their requirement for a robust BI solution to harness the vast amount of data available on their daily refreshed servers for updated analysis and faster decision-making.

Firstly, we designed the dimensional model. This involved understanding the context and determining the most critical measures, enabling us to create the dimensions and the fact table.

Subsequently, we implemented that design in Microsoft Fabric, starting with the creation of a Data Lakehouse and Data Warehouse, the various dimensions conceptualized, and finally, the creation of pipelines to ensure everything runs smoothly.

Then, Power BI reports were created, divided into six different categories: 2023 Overview, Performance Analysis and Forecast, Driver Analysis, Client Analysis, Location Analysis, and Operational & Car Analysis, facilitating narrative construction and assisting in monitoring business performance. The reports were developed with a simple design and a color palette that reflects the client's identity, predominantly shades of blue, green, and white.

Finally, the solution met our and the customer's expectations, as it addresses the company's business needs. It also provides them with the ability to understand their current business position and the opportunity to forecast future scenarios. We managed to answer questions by organizing key themes to facilitate information storytelling, always keeping the client's image in mind, that aligns with Ride4ALL's identity, thereby maintaining a visual identity that makes the entire project more personalized and tailored to the needs of Ride4ALL.

11 Appendix: Executed procedures

In this appendix section we have the recently completed tasks. Also includes the 2 support tables we refer to above.

1. Duplicate\_Locations\_to\_map

This Ride4All\_VALIDATIONSTG\_2024 dinamic table is used to map the supplicates directly in the Dataflo\_Fact\_Rides.

Data preview							Showing 6 rows	Search
	12L id	ABC City	ABC Country	ABC District	ABC Is_Active	12L ActiveID		
1	18	Beja	Portugal	Beja	YES	18		
2	20	Beja	Portugal	Beja	NO	18		
3	45	São Pedro do Sul	Portugal	Viseu	YES	45		
4	101	São João da Madeira	Portugal	Aveiro	YES	101		
5	102	São João da Madeira	Portugal	Aveiro	NO	101		
6	105	São Pedro do Sul	Portugal	Viseu	NO	45		

Figure 31. Duplicate\_Locations\_to\_map table.

2. Data Warehouse Loading and validation rules with email notification.

Pipeline run ID: dc32094c-b1fe-48d4-848b-8210667f0be3

Pipeline status ✔ Succeeded

Showing 1 - 21 items

Activity name	Activity status	Run start	Duration
Successful DW Loading and Validation	<span>✔ Succeeded</span>	5/5/2024, 3:19:48 PM	Less than 1s
Validation Rules	<span>✔ Succeeded</span>	5/5/2024, 3:18:06 PM	1m 41s
Dataflow Fact Rides	<span>✔ Succeeded</span>	5/5/2024, 3:14:32 PM	3m 33s
Wait until Dim Dataflows are complete	<span>✔ Succeeded</span>	5/5/2024, 3:14:30 PM	2s
Dataflow_Date	<span>✔ Succeeded</span>	5/5/2024, 3:12:45 PM	1m 3s
Dataflow_Currency	<span>✔ Succeeded</span>	5/5/2024, 3:12:45 PM	1m 3s
Dataflow_Time	<span>✔ Succeeded</span>	5/5/2024, 3:12:45 PM	1m 3s
Dataflow_Car	<span>✔ Succeeded</span>	5/5/2024, 3:12:45 PM	1m 3s
Dataflow_Location	<span>✔ Succeeded</span>	5/5/2024, 3:12:45 PM	1m 34s
Dataflow_Drivers	<span>✔ Succeeded</span>	5/5/2024, 3:12:45 PM	1m 43s
Dataflow_Client	<span>✔ Succeeded</span>	5/5/2024, 3:12:45 PM	1m 3s
Wait until the fact is clear	<span>✔ Succeeded</span>	5/5/2024, 3:12:43 PM	2s
SQL clear Rides facts in DW	<span>✔ Succeeded</span>	5/5/2024, 3:12:26 PM	16s
Wait until all dimensions are clear	<span>✔ Succeeded</span>	5/5/2024, 3:12:24 PM	2s
SQL clear dim Time in DW	<span>✔ Succeeded</span>	5/5/2024, 3:12:07 PM	16s
SQL clear dim Car in DW	<span>✔ Succeeded</span>	5/5/2024, 3:12:07 PM	15s
SQL clear dim Client in DW	<span>✔ Succeeded</span>	5/5/2024, 3:12:07 PM	15s
SQL clear dim Location in DW	<span>✔ Succeeded</span>	5/5/2024, 3:12:07 PM	15s
SQL clear dim Driver in DW	<span>✔ Succeeded</span>	5/5/2024, 3:12:07 PM	16s
SQL clear dim Currency in DW	<span>✔ Succeeded</span>	5/5/2024, 3:12:07 PM	15s
SQL clear dim End_Date in DW	<span>✔ Succeeded</span>	5/5/2024, 3:12:07 PM	15s

Nova Information Management School, Lisbon, Portugal

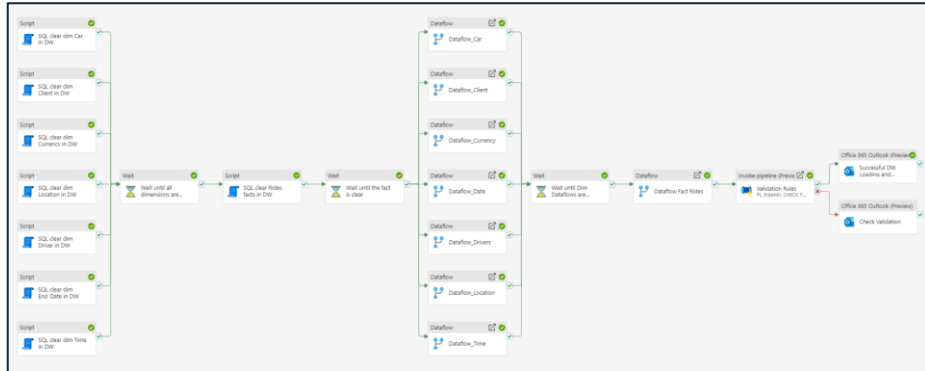


Figure 32. Successful pipeline running

### 3. log\_quality\_checks: Validation Rules and Row Count.

Warehouses	id_check	ABC table_name	ABC etl_checktype	ABC description_result	ABC etl_result
1	10	Dim_Client	get total row count for each table	total row count: 10000	NA
2	10	Dim_Car	get total row count for each table	total row count: 500	NA
3	10	Dim_Driver	get total row count for each table	total row count: 1000	NA
4	10	Dim_Location	get total row count for each table	total row count: 130	NA
5	10	Dim_Date	get total row count for each table	total row count: 1461	NA
6	10	Dim_Time	get total row count for each table	total row count: 1440	NA
7	10	Dim_Currency	get total row count for each table	total row count: 1	NA
8	10	Fact_Rides	get total row count for each table	total row count: 1048572	NA
9	7	Fact_Rides	check foreign key relationships in fact table	number of invalid foreign key relationships: 0	OK
10	6	Fact_Rides	check uniqueness of all fact attributes	number of rows NOT unique: 0	OK
11	3	Dim_Currency	check uniqueness of all dim attributes	number of rows NOT unique: 0	OK
12	5	Dim_Location	check uniqueness of all dim attributes	number of rows NOT unique: 0	OK
13	1	Dim_Car	check uniqueness of all dim attributes	number of rows NOT unique: 0	OK
14	4	Dim_Driver	check uniqueness of all dim attributes	number of rows NOT unique: 0	OK
15	2	Dim_Client	check uniqueness of all dim attributes	number of rows NOT unique: 0	OK

Figure 33. Results from the rules

### 4. Email Notification:

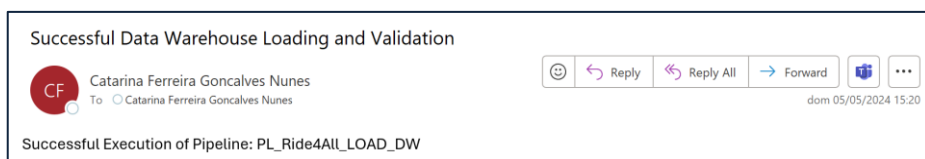
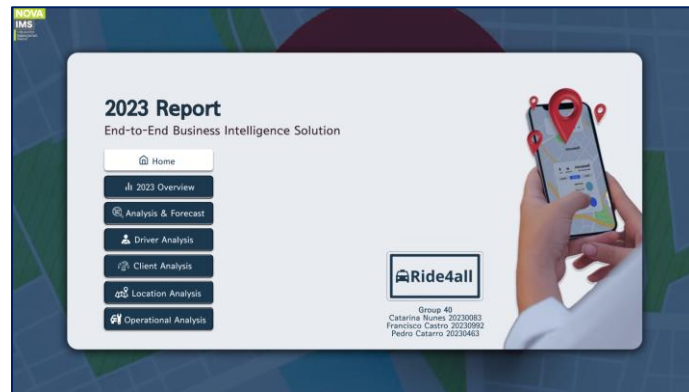


Figure 34. Email of Successful pipeline running

*Nova Information Management School, Lisbon, Portugal*

## 5. Power BI Report



*Figure 35. Report Cover*

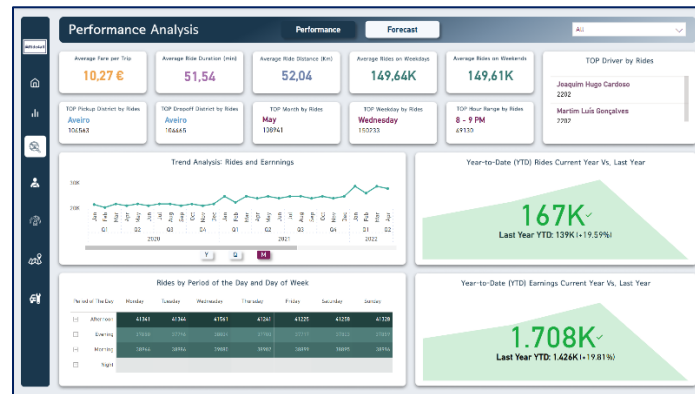
## 2023 Overview



*Figure 36. 2023 Overview*

*Nova Information Management School, Lisbon, Portugal*

## Performance Analysis & Forecast



*Figure 37. Performance Analysis*



*Figure 38. Forecast*

Nova Information Management School, Lisbon, Portugal

## Driver Analysis

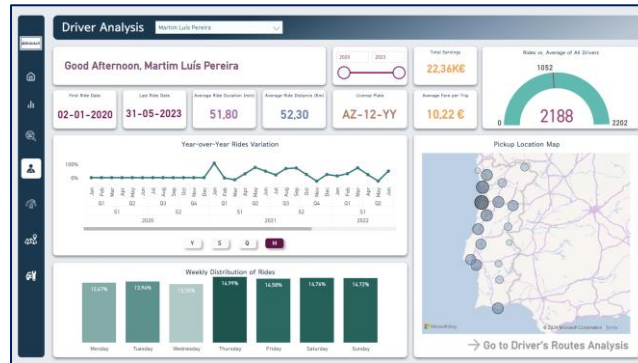


Figure 39. Driver Profile

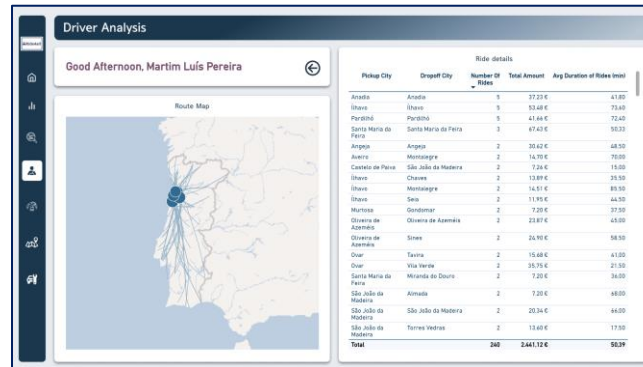


Figure 40. Driver's Routes Analysis

## Client Analysis



Figure 41. Client Profile



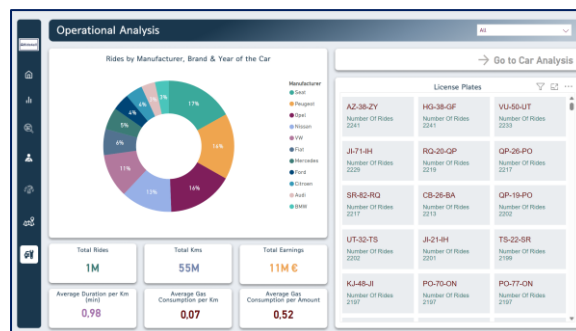
*Nova Information Management School, Lisbon, Portugal*

## Location Analysis

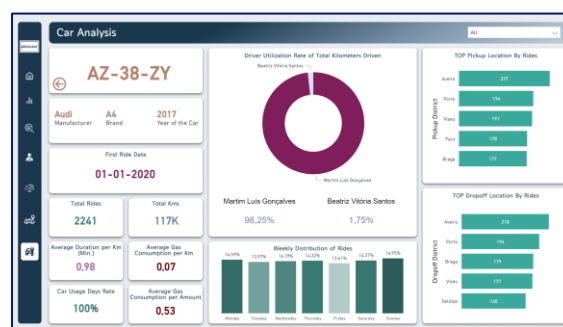


*Figure 42. Location Analysis*

## Operational Analysis & Car Analysis



*Figure 43. Operational Analysis*



*Figure 44. Car Analysis*